

An Examination of the Improved Wald Test for Differential Item Functioning Detection
with Multiple Groups

By

© 2015

Hsiang-Feng Chen Carroll

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of
the University of Kansas in partial fulfillment of the requirements for the
degree of Doctor of Philosophy.

Chairperson: John Poggio, Ph.D.

Meagan Patterson, Ph.D.

Vicki Peyton, Ph.D.

William Skorupski. Ed.D.

Wei Wu, Ph.D.

Date Defended: September 4, 2015

The Dissertation Committee for Hsiang-Feng Chen Carroll
certifies that this is the approved version of the following dissertation:

An Examination of the Improved Wald Test for Differential Item Functioning Detection
with Multiple Groups

Chairperson: John Poggio, Ph.D.

Date approved: September 4, 2015

Abstract

Methods for identifying items that function differentially (DIF) across groups have over time become increasingly essential for establishing validity in psychometric testing. There is a vital need for simultaneously comparing multiple groups, evidenced by the widespread proliferation of international assessments in educational testing. Historically, DIF methods have been formulated to only address pairwise group comparisons; methods that address the two or more groups' case in a latent variable context are a comparatively recent phenomenon. This study evaluates the effectiveness of Wald-1, a newly developed DIF detection approach for multiple group comparisons. Data were simulated under the three-parameter logistic (3PL) Item Response Theory (IRT) model, with the explicit design of simulation conditions and parameters guided by empirical data. Results were examined in the context of statistical power and Type I error rate under various combinations of experimental conditions: (a) the number of test groups, (b) the number of candidate items with DIF, and (c) the number of anchor items with DIF.

The results indicate that Wald-1 performs well in the identification of DIF items, collapsing across all other variables such as the number of groups, conditional on the existence of a DIF-free anchor set. The effectiveness of Wald-1, practical implications of the results, and considerations of future research are examined and discussed thoroughly. Reliable and effective anchor selection methods are prerequisite of excellent performance for Wald-1 detecting DIF.

Acknowledgements

My journey toward obtaining this doctorate has coincided with several significant life events: marriage, entry into parenthood, and acceptance into the first full-time position for my career. Looking back at this journey, I would like to express my gratitude to many people for helping and encouraging me to complete my dissertation.

I would like to dedicate this dissertation to my husband, Ian Carroll, and our first child Marcus. Often I asked myself, how small was the chance that two people from different countries, with me traveling to the opposite side of the Earth, getting to know each other would lead to marriage? During my years at the University of Kansas, I was fortunate to have met and fallen in love with my life partner and best friend Ian. Since the day we have been together, he has completely supported my dreams and my goals. He has inspired me to achieve more than I thought possible. After all of the countless nights we have spent editing my innumerable drafts, troubleshooting code, and discussing how to present analysis results, he has always been my pillar of strength through all my ups and downs. Words cannot describe his selflessness and my gratitude. Ian, thank you for being my rock and my safe place to land.

Despite a very large distance and leaving home, my father Jau-Jiun Chen and my mother Tzu- Jung Yen have always supported my dreams and made them possible in so many ways. For my entire life they have only wanted what was best for me, even if that means living far away. I wouldn't have the courage to take many big risks, such as coming to the United States to pursue my education, without them standing behind me.

I would never have been able to finish the program and dissertation without the guidance and support from my academic advisor and life mentor, Dr. John Poggio. He has always been

there for me -- he supported me throughout my down days and celebrated my achievements with me along the way.

My dissertation committee was instrumental in the process. I would like to thank Dr. Meagan Patterson for her continuous encouragement and timely support throughout my Ph.D. study. Dr. William Skorupski always provided me with invaluable feedback and stimulating commentary; I am very grateful to have had the opportunity to learn from him. A big thank you to Dr. Wei Wu for her willingness to participate in my defense committee at the last moment without any hesitation; she was generous with time and promptly replied to e-mails. I would also like to thank Dr. Vicki Peyton for guiding my research for the past several years. Additionally, I wish to thank Dr. Carol Woods and Dr. Michelle Langer for inspiring me to study on this topic and providing knowledgeable input.

Many friends and colleagues have supported me (emotionally and intellectually) during the entire dissertation process: Betty and Dale Guder, Jared Harpole, Mian Wang, Richard Kinai, Leslie Shaw, Dr. Barbara Wells, Dr. Chi-Hsun (Jessie) Chiu, Shonda Anderson, Peter Griggs, Dr. Katherine Hole, Dr. Aki Sato, Dr. Megha Ramaswamy, and Summer Cheng. Finally, my sincere and special thanks go to my father-in-law Dr. John Carroll, who treats me like his own daughter and always supports me toward accomplishing my goals.

謹以本論文，獻給我最親愛的父親陳昭鈞先生和母親嚴紫蓉女士。遊美負笈求學十載，在太平洋另一端的他們，一路給予我最堅實的支持與力量。而今自己成為母親之後，方才真切體會父母在我成長過程中的無盡付出與親恩浩瀚。若我有一點小小的成績，那是因為您們讓我想要蛻變成為更美好的自己。謝謝您們永遠給我溫暖厚實的肩膀，守護著逐夢的我。十年磨劍，終告功成。我期許自己成為您們和家人朋友的驕傲。

Table of Contents

Abstract	iii
Acknowledgements	iv
Table of Contents	vi
List of Tables	viii
List of Equations	ix
List of Figures	x
Chapter I: Introduction.....	1
General Problem	2
Purpose Statement	9
Research Questions.....	9
Definitions of Variables	10
Summary.....	10
Chapter II: Literature Review.....	12
Brief Statement about Item Response Theory	12
History and Perspectives on Differential Item Functioning Procedures	13
Differential Item Functioning Detection Methods Extended to Multiple Groups.....	15
The Mantel-Haenszel Procedure.....	15
Q_j Statistic	20
Multiple Indicator Multiple Cause (MIMIC) Models.....	22
IRT Likelihood Ratio Tests	24
Improved Wald Tests.....	25
Chapter III: Methodology.....	30

Empirical Parameter Distributions	30
Design	33
Outcomes	36
Chapter IV: Results	37
The Relation between Number of Groups, Anchor Contamination, and Type I Error Rate ..	37
A New Factor: Proportion of Sample with DIF.....	42
Group Mean Effects.....	44
ANOVA Results	46
Conclusion	47
Chapter V: Discussion and Conclusions	49
Conclusions.....	52
Limitations and Future Directions for Research.....	54
References.....	55
Appendix A: Three-Way ANOVA Results	63

List of Tables

Table 1. The 2 (groups)-by-2 (item scores)-by-M (score level) contingency Table Viewed in 2x2 Slices	16
Table 2. Definition of the DIF Categories A, B, and C Based on the MH D-DIF Statistic	19
Table 3. Factors Used in Simulation Study	28
Table 4. Summary Statistics for Math Proficiency	32
Table 5. Group Conditions of the Reference and Focal Groups	34
Table 6. Type I Error Rate by Anchor Contamination	38
Table 7. Summary table of Type I error rates for Figure 10	41

List of Equations

(1) IRT 3PL model.....	13
(2) Null DIF hypothesis for the Mantel-Haenszel method	16
(3) Constant odds ratio hypothesis of the Mantel-Haenszel method	17
(4) Mantel — Haenszel chi-square statistic	17
(5) Estimated constant odds ratio of Mantel — Haenszel method	18
(6) Mantel — Haenszel odds ratio in the delta metric	18
(7) IRT Likelihood Ratio Tests.....	25
(8) Lord's DIF significance test	25

List of Figures

Figure 1. Item Characteristic Curve: Uniform DIF	3
Figure 2. Item Characteristic Curve: Non-Uniform DIF	3
Figure 3. Basic MIMIC Model for DIF detecting.....	23
Figure 4. Ability parameters distribution of each simulated group.	32
Figure 5. Simulation study: pools of anchor and candidate items.	33
Figure 6. Illustration of the simulation design.	35
Figure 7. Box-and-whisker plot of Type I error rates by anchor contamination: three-group conditions.....	39
Figure 8. Box-and-whisker plot of Type I error rates by anchor contamination: six-group conditions.....	39
Figure 9. Side-by-side comparisons of Type I error rates for three and six groups.	40
Figure 10. Type I error rates with only one DIF group: three-group versus six-group conditions.	41
Figure 11. Type I error rates by percentage of overall sample with DIF and anchor contamination.....	43
Figure 12. Statistical power by percentage of sample with DIF and anchor contamination.	44
Figure 13. Type I error rate for 50% contamination, one group with DIF in six-group conditions, equal sample size.	45
Figure 14. Interaction plot: percentage of samples with DIF by anchor contamination.....	63
Figure 15. Interaction plot: anchor contamination by number of groups	64
Figure 16. Interaction plot: percentage of the overall sample with DIF by number of groups	65
Figure 17. Three-way interaction plot (three-group conditions)	66

Figure 18. Three-way interaction plot (six-group conditions).....	66
---	----

Chapter I: Introduction

Do you know how many different primary languages are spoken on the University of Kansas (KU) campus? As of the fall 2014 semester, over 40 languages were spoken on campus, with international students at KU hailing from 101 countries. This fact is fascinating when you consider the far-ranging diversity in both language and culture for students at this institution.

Imagine that the University needs to implement a mathematics exam for incoming freshmen who have varying linguistic and cultural backgrounds, and that the test results will determine whether or not the students are qualified for different majors. In this case, undoubtedly, both the University and the examinees expect that test results will fairly and equitably reflect the students' achievement. However, numerous studies have provided evidence suggesting that test scores are not necessarily trustworthy for the purpose of representing individuals' traits due to measurement bias (Stark, Cherynshenko, & Drasgow, 2004). Specifically, people with different cultural backgrounds may perceive the same test item in different ways, which in turn affects the probability of answering the question correctly, even if they have equivalent ability levels. Another example could be that, in a state-wide assessment for 8th graders, a question (or a group of questions) set in the context of dirt bike racing or duck hunting would not be as readily accessible to girls as to boys. Therefore, it would result in an unfair advantage for boys and an unfair disadvantage for girls.

These two examples reveal how existing heterogeneity among students challenges researchers in the field of educational measurement. In order to ensure the validity and fairness of measurements in students' educational achievement across cultural backgrounds and demographics, testing professionals have developed methods and refined existing psychometric techniques to identify items that function differentially across distinct groups, such as by age,

race, culture/language, socioeconomic status (SES), and mobility.

General Problem

When individuals from different groups (gender, majority/minority, SES, etc.) perform differently on a test item, it can be the result of a true difference between the groups regarding the construct being evaluated, or the item itself might be causing a difference to appear, regardless of whether a group-level difference exists. This difference in the item score, above and beyond group differences on the construct, is referred to as *differential item functioning* (DIF). Broadly defined, DIF refers to the presence of differences in individual item characteristics across groups, which can be observed graphically as differently sloped or horizontally shifted item characteristic curves (ICCs) when the item parameters for each group are plotted (Lord, 1980; Thissen, Steinberg, & Wainer, 1988). These differences in item parameters are thought to indicate empirical differences in an item's tendency to accurately or inaccurately estimate an examinee's standing on a latent trait (often ability) among different groups. Thus, such differences may be sources of bias in the form of unfair advantages for some groups and unfair disadvantages for other groups. Applied methods for investigation of the statistical and substantive sources of such differences in item functioning across groups have become increasingly relevant tools for improving the validity of assessments in education, medicine, psychology, employment and hiring, career advancement, and other fields.

There are two types of DIF: uniform and non-uniform DIF. Uniform DIF means that one group always has advantage in answering a test item correctly at all ability level. Figure 1 demonstrates that almost all members of group 2 are favored over almost all members in group 1 when responding to a test item correctly. For instance, two test takers are at the same ability level ($\theta = 0$) but belong to different groups. An examinee of group 2 has an approximate 0.84 chance of

correctly responding to the question, but another examinee of group 1 has only 0.62 chance of correctly responding to the same item. In the case of non-uniform DIF, one group is at an advantage up to a proficiency/ability level and from that point on this group is at a disadvantage when responding to a question correctly. Figure 2 illustrates an example of non-uniform DIF. Two item characteristic curves intersect around ability level of -0.4. In this example, up to ability level of -0.4, members of groups 3 have advantage in answering this question correctly. Beyond this intercept (i.e. $\theta > -0.4$), the situation is reversed and group 4 is favored instead.

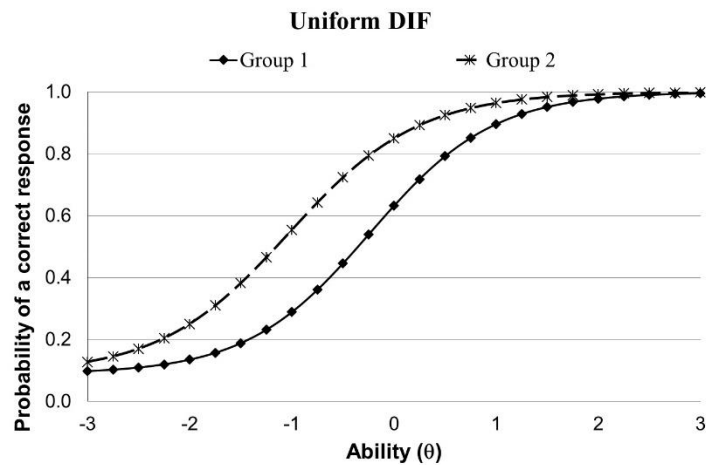


Figure 1. Item Characteristic Curve: Uniform DIF

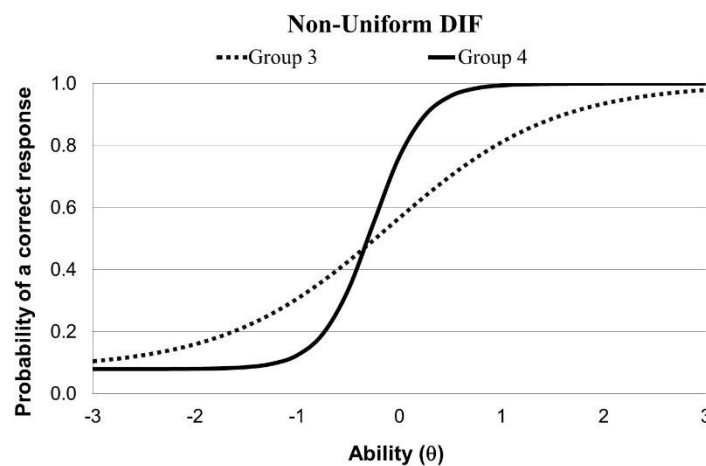


Figure 2. Item Characteristic Curve: Non-Uniform DIF

DIF detection approaches are often classified into two main categories: (1) the observed summed scores approach and (2) the latent variable approach. The observed-score method has been popularly investigated since the 1970s, especially the Mantel-Haenszel (M-H) procedure (Applebee, Langer, & Mullis, 1987; Holland & Thayer, 1988; Zwick, 2012; Zwick & Ercikan, 2005). In addition to the M-H procedure, logistic regression (Swaminathan & Rogers, 1990), SIBTEST (Shealy & Stout, 1993), and Quade's Family of Average Conditional Ordinal Measures (Woods, 2009) were also included in the domain observed-score approach.

The latent variable approach has received much attention in the past decade, and numerous studies have tried to develop improved versions of the original approaches, including the item response theory (IRT) approach. Latent variable methods include Lord's (1977, 1980) Wald (1943) χ^2 test, multiple indicator multiple cause (MIMIC) models (Finch, 2005; Jöreskog & Goldberger, 1975; Muthe'n, 1985, 1989; Woods, 2009), the MIMIC-interaction model (Woods & Grimm, 2011), and the IRT likelihood ratio tests, which is often called IRT-LR-DIF (Thissen et al., 1988).

Regardless of the category into which a particular method falls, most DIF detection methods utilize a reference group and a focal group to specify a comparison. In other words, many of the methods stated above share the limitation that item functioning can be compared across two groups only, which restricts the analytic approach to pairwise comparisons. However, there is a vital need for simultaneously comparing multiple groups because of the need to decrease the type I error rate and time-efficiency (Ellis & Kimmel, 1992; Kim, Cohen, & Park, 1995; Schmitt & Dorans, 1990). For example, as it is known for some methods (e.g., multiple *t*-tests), type I error rate inflation is less likely to occur when the number of comparisons is rather small (Penfield, 2001). Furthermore, among international assessments, the participating test-

takers represent a rich diversity of languages, cultures, and customs. It is important to effectively and efficiently flag test items that perform differentially between groups. A valid multiple-group comparison is theoretically more efficient than all possible or a reduced set of paired comparisons. At present, a few DIF detection procedures have implied DIF detection among multiple groups. The methods that allow more than two groups for simultaneous comparison include the M-H procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) and generalized Mantel–Haenszel method (GMH; Somes, 1986; Penfield, 2001; Zwick, Donoghue & Grima, 1993); the Q_j statistic, an extension of Lord’s Wald χ^2 test (Kim, Cohen, & Park, 1995); the MIMIC models (Finch, 2005; Jöreskog & Goldberger, 1975; Muthe’n, 1985, 1989; Woods, 2009; Woods & Grimm, 2011); IRT likelihood ratio tests (Thissen et al., 1988); and the improved Wald tests (Cai, Yang & Hansen, 2011; Langer, 2008; Woods, Cai & Wang, 2012). However, previous literature has reported that many of the methods listed above, such as the M-H procedure and the Q_j statistic test, have their own limitations. More details regarding each of the five above-mentioned methods will be discussed in Chapter II.

The M-H procedure could detect DIF in more than two groups when performed by Penfield (2001), yet a major limitation is its inability to detect non-uniform DIF. Furthermore, the M-H procedure has difficulty in differentiating the mean variances in true ability among groups from the existence of DIF. This issue, in other words, is often referred to as the “incorrect identification of DIF” (Finch, 2005, p. 291). When employing the M-H procedure, Finch (2005) concluded that the error rate inflated when the sample size of a focal group was larger and the group means between the groups were very close. This is a universal problem inherent in all observed summed-score approaches because the control variable is the summed score rather than a latent variable.

Kim, Cohen, and Park (1995) extended Lord's original approach by developing the Q_j statistic. Evidence suggests that the Q_j statistic may effectively identify DIF between three groups by comparing item parameter vectors, but further investigation asserted that this study suffered from several limitations (Langer, 2008). First, the authors failed to consider "the density of examinees in the sample along the ability continuum (Langer, 2008, p.34)", and may have falsely detected test items with DIF and inflated Type I error rates. In addition, the information given for the item parameters' error covariance matrix was inadequate for the record.

More recently, Langer (2008) proposed an alternative of the two-stage procedure for improving the original Wald test. Lord's Wald test was originally designed to detect the difference in b parameters between two groups, which necessarily means this approach was only able to detect uniform DIF at first. Later, Lord created a more generalized version of the test that incorporated information about a parameters in order to detect non-uniform DIF. More specifically, the Wald statistic incorporates information from the covariance matrix of the differences of item parameter estimates between groups, along with the actual values of these differences themselves, to create a chi-square statistic.

Langer's (2008) contributions to this were twofold. First, Langer made a change in the linking/equating procedure, which is an essential step before DIF detection takes place, especially when researchers have data consisting of various test forms and different groups of examinees. The purpose of the linking/equating procedure is to place the item parameters, as well as ability estimates from the reference and focal groups (or two forms of an examination), on the same metric so that the scores can be compared and used interchangeably. Langer made a change in the linking/equating procedure from the Stocking-Lord (1983) approach to concurrent

calibration (Kolen & Brennan, 2004; Langer, 2008, p. 10). One of the main differences between the Stocking-Lord (1983) approach and concurrent calibration is that the Stocking-Lord approach obtains the estimate item parameters from different test forms through separate runs, yet concurrent calibration is able to simultaneously estimate all item parameters from various test forms in a single run (Kim & Cohen, 2002). Several studies have found that concurrent calibration outperforms the Stocking-Lord (1983) approach in the accuracy of estimated scores and the stability of recovering item parameters (Kim & Cohen, 2002; Petersen, Cook, & Stocking, 1983; Tian, 2011).

In addition to the change in the linking/equating procedure, the improved Wald test (Cai, et al., 2011; Langer, 2008) was demonstrated to have better performance than Lord's original approach because the covariance matrix is estimated using the supplemented expectation maximization (SEM) algorithm (Cai, 2008; Meng & Rubin, 1991). The statistically desirable attributes of the improved Wald test are as follows. First, the improved Wald test is currently considered a practical alternative to IRT-LR-DIF; the latter has been shown to be the most powerful method for DIF detection in several previous studies (Thissen, et al. 1993; Wainer, 1995). More specifically, Langer specified that the improved Wald test is "asymptotically equivalent to the likelihood ratio test (Thissen, et al. 1993) and less computationally-intensive" (2008, p.2). That means the procedure of employing this approach to DIF detection across several groups simultaneously would take much less computation time. Furthermore, when comparing to GMH method and basic MIMIC model, the improved Wald tests are considered superior because its utilities in detecting not only uniform DIF, but also non-uniform DIF (Woods et al., 2012). Although an improved variant of the MIMIC DIF model (i.e., MIMIC-interaction model) has been created to detect non-uniform DIF (Woods & Grimm, 2011), the

method has not yet been implemented fully and correctly in popular software programs (i.e. Mplus, R). The improved Wald tests, by contrast, have the advantage in the ease of multiple group comparisons by means of IRTPro and flexMIRT (Cai, 2012). Nevertheless, this method may suffer from some limitations. For instance, when comparing it to approaches such as the M-H procedure, the Langer-improved Wald test requires more assumptions and a larger sample size. Additionally, the Wald approach theoretically could have covariates included in the model so that other factors (e.g., demographic characteristics), except theta, could be statistically controlled. However, candidate software such as IRTPro and flexMIRT currently does not allow specification of covariates.

Over the years, researchers have established the pros and cons of the M-H procedure and Lord's Wald test. However, the precise limitations of the Langer-improved Wald test are still unclear. The Langer-improved Wald test was developed fairly recently; thus, very little literature has evaluated this new approach. Amongst the few that have evaluated the Langer-improved Wald test, Woods, Cai, and Wang (2012) assessed Langer's Wald test via simulation by employing two different equating algorithms, Wald-1 and Wald-2. Wald-1 detected DIF with designated anchor items to assess candidate items and estimate group differences. In contrast, Wald-2 did not require designated anchor items because this approach treated all items as anchor items ("test all items, anchor all items"), which was how Langer (2008) detected DIF. The results indicate that Wald-1 (test certain anchor items) outperforms Wald-2 (using all items as anchors) in important metrics like correctly estimating mean differences on theta and their variability.

Woods et al. (2012) examined Langer's method and found that, under an assumption of using accurate anchor items for DIF detection, Wald-1 results in better performance compared to Wald-2, which was Langer's original strategy. A logical follow-up to this study would be a

comparative evaluation of method performance when designated anchors are not correctly specified in full.

Furthermore, Langer (2008) and Woods et al. (2012) implemented the improved approach for simultaneous comparisons among no more than four groups in their studies. It may be beneficial to the field and future research to explore precisely how many groups the Wald-1 test can simultaneously evaluate while still maintaining an appropriate level of statistical accuracy under varying experimental conditions.

Purpose Statement

The purpose of this study is to examine the statistical properties of Wald-1 DIF testing by means of simulated data. This study has two purposes. The first is evaluating the performance of Wald-1 with 3 and 6 groups. The second is measuring the impact on test performance when anchors are inaccurately specified to varying degrees of severity. Anchor misspecification must be examined if Wald-1 DIF testing is to be rigorously evaluated, not merely because the method makes the assumption of having a pure anchor set, but also to provide a necessary realism to the simulation.

Research Questions

I am seeking to answer two research questions:

- 1) Wald-1 requires anchors and gives results similar to those of IRT-LR-DIF. Among the comparisons between four groups, Wald-1 is more computationally efficient, expedient, and convenient than IRT-LR-DIF. How does Wald-1 perform when detecting DIF with more the two groups simultaneously?
- 2) Suppose Wald-1 uses improperly specified anchor items to detect DIF. How would this impact test performance?

Definitions of Variables

In the present DIF study, the performance of focal groups on each multiple choice mathematic item will be compared with the performance of the reference group (i.e., U.S. students), after controlling for ability differences between countries. Below are the definitions of the terms used in this study:

- 1) **Differential item functioning (DIF):** An item functions differently when people from different manifest groups (e.g., males and females) do not have equal probability of a correct answer, even if they have the same level of ability (de Ayala, 2009).
- 2) **Reference group:** Reference groups, sometimes referred as *majority groups*, are compared to the focal group (de Ayala, 2009). In this study, as stated previously, students in the U.S. comprise the reference group.
- 3) **Focal group:** The focal group is the believed-to-be-disadvantaged group being studied (de Ayala, 2009). In this study, non-U.S. countries comprise the focal groups.
- 4) **Anchor items:** Anchors refer to a common set of items (that are DIF-free) used to place the reference group and focal groups on the same metric.

Summary

The application of detecting items that function differently across groups has become an increasingly relevant tool for improving assessment validity. Most DIF detection methods share the limitation of pairwise comparisons, though there is a vital need for comparing multiple groups simultaneously because of time efficiency and the need to decrease the type I error rate. This study aims to evaluate the effectiveness of a newly developed DIF detection approach, called Wald-1, which is mean to “test candidate items for DIF with designated anchor items” (Houts & Cai, 2012, p. 40) under varying experimental conditions. The purpose of this research

is twofold: (1) assess the performance of Wald-1 with more than three groups and (2) evaluate the impact on test performance when anchor items are not correctly specified in full. Simulated data will be used to examine the statistical properties of this new approach.

Chapter II: Literature Review

The focus of this present study is to examine the performance of the Wald-1 method on identifying items that function differently (often referred to as Differential Item Functioning, or DIF) between groups in various conditions, after controlling for their ability or proficiency. This chapter will first provide an introduction to Item Response Theory (IRT) and its implementation in DIF detection, followed by a brief review of the progression of DIF methods, and finally some common DIF methods applicable to comparisons across more than two groups simultaneously.

Brief Statement about Item Response Theory

In many fields such as education and psychology, it is common to develop instruments with the intention of measuring a variable of interest (e.g., intelligence, anxiety, mathematical ability) that cannot be measured directly like weight, height, or temperature. Such variables are often referred to as *traits* (Baker, 2001). Item Response Theory (IRT) develops non-linear regression lines, called item characteristic curves (ICCs), to estimate how much of a trait an individual possesses. ICCs allow researchers to visualize examinee ability (θ) in relation to the probability of a correct response on a specific test item (Lord, 1983).

IRT refers to “a class of models describing the relationship between individual item responses and the construct(s) measured by the test (Thissen, et al. 1993, p.67)”. Some well-known assumptions of IRT include unidimensionality, local independence, and monotonicity. Unidimensionality implies that only one construct is being measured. Local independence is, in relation to unidimensionality, the assumption that the trait is the only variable influencing the probability of a correct answer. Item responses are not correlated with each other after controlling for ability. Individual’s responses to test items are also uncorrelated with one another. Once examinee ability (θ) is known, one can predict the chance that the subject will provide a

correct answer. Monotonicity means that item response is monotonically related to the individual's ability. One person with higher ability has a better chance to answer an item correctly than another with lower ability (Kaplan, 2004). Some extensions of traditional IRT models (a) allow one or more traits, (b) use discrete or continuous data, and (c) use data scored dichotomously or polytomously. The commonly used models (e.g. 1-parameter, 2-parameter, and 3-parameter logistic models) include the following elements: examinee trait (θ), a-parameter (discrimination), b-parameter (difficulty), and c-parameter (guessing) (Lord, 1983).

This present study focuses on the 3-parameter logistic (3PL) model. Equation 1 shows a 3PL model for the probability of person i answering item j correctly.

$$P(U_{ij} = 1 | \theta_i) = c_j + (1 - c_j) \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}}, \quad (1)$$

where D equals 1.702 (a scaling coefficient that connects logit and probit link functions), θ_i is examinee i 's ability, a_j is the discrimination parameter for item j , b_j is the difficulty parameter for item j , and c_j refers to the pseudo-guessing parameter for item j (Thissen & Wainer, 1982).

History and Perspectives on Differential Item Functioning Procedures

One of the earliest studies investigating the items that function differently between groups was conducted in 1964 by Cardall and Coffman. In this study, those items being detected were called the *item bias* in a test (Angoff, 1993; Cardall & Coffman, 1964); the authors employed ANOVA to compare groups of black examinees with white examinees in SAT administration during 1963. Their research opened a series of subsequent studies evaluating the interaction of items with different groups (Mellenbergh, 1989), such as language (Angoff & Sharon, 1974), race, and socioeconomics (Cleary & Hilton, 1968). Even so, this approach was not widely endorsed by the majority of the psychometric society at the time (Angoff, 1993).

An enormously influential breakthrough in the course of test development theory took place in the dissertation of Frederic M. Lord (1952), as well as in a relevant study published by Lord and Novick (1968) in which the IRT model was first introduced. Soon after, the concept of IRT became gradually popular after the late 1970s. Along with the advent of computers and the corresponding accessible power of technology, IRT was found to be beneficial to DIF studies because DIF theory was built upon the foundation of IRT (Angoff, 1993).

The concept with respect to DIF, based upon IRT, was specified early in these two pioneering publications. First, Lord (1980) said,

If each test item in a test had exactly the same item response function in every group, then people of the same ability or skill would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased. If on the other hand, an item has a different item response function for one group than for another, it is clear that the item is biased. (p. 212)

Seven years later, Thissen (1987) provided his perspective of DIF in a conference discussion:

... an expression which describes a serious threat to the validity of tests used to measure the aptitude of members of different populations or groups. Some test items may simply perform differently for examinees drawn from one group or another or they may measure “different things” for members of one group as opposed to members of another. Tests comparing such items may have reduced validity for between-group comparison, because their scores may be indicative of a variety of attributes other than those the test is intended to measure. (p. 1)

In order to evaluate the probability of how individuals with equivalent abilities would answer a given question correctly, IRT is utilized to observe S-shaped traces, which are now

well-known as ICCs (Angoff, 1993, p. 8). An ICC is composed of three distinct parameters that measure the discriminating ability of the item (a -parameter); the difficulty of the item (b -parameter); and the pseudo-guessing parameter of the item (c -parameter), which “reflects the nonzero probability of a low trait examinee obtaining a correct answer by chance” (Reise & Waller, 2003, p. 165). It appears that the model employing three item parameters (a , b , c) is exhaustive because differences in item difficulty, discriminating power, and guessing between groups are taken into consideration (Angoff, 1993).

Differential Item Functioning Detection Methods Extended to Multiple Groups

A variety of methods are available to detect items that function differently between groups, but only a few are capable of conducting simultaneous DIF detection between multiple groups. This section will introduce five methods presently employed to flag items as differentially functioning: the M-H procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) and generalized Mantel–Haenszel method (GMH; Somes, 1986; Penfield, 2001; Zwick et al., 1993); the Q_j statistic, an extension of Lord’s Wald χ^2 test (Kim, Cohen, & Park, 1995); the MIMIC models (Finch, 2005; Jöreskog & Goldberger, 1975; Muthe’n, 1985, 1989; Woods, 2009; Woods & Grimm, 2011); IRT likelihood ratio tests (Thissen et al., 1988); and the improved Wald tests (Cai et al., 2011; Langer, 2008; Woods et al., 2012). This section describes these five methods in some detail.

The Mantel-Haenszel Procedure

Mantel and Haenszel (1959) first introduced a new approach for use in the field of epidemiology. Later, Holland and Thayer (1988) adapted this method to detect DIF items. The basic concept of the M-H procedure is to employ a set of M 2-by-2 contingency tables, where M is the number of score categories on the matching variable, such as the total summed score

(Zwick, 2012). The contingency tables are composed of the group membership (i.e., the reference group and focal group) and the item score (right and wrong answer). Table 1 provides an example of one 2-by-2 contingency table, which was originally mentioned by Dorans and Holland (1993, p. 39):

Table 1

The 2 (groups)-by-2 (item scores)-by- M (score level) contingency Table Viewed in 2x2 Slices

Group	Item Score		
	Right	Wrong	Total
Focal group (f)	R_{fm}	W_{fm}	N_{fm}
Reference group (r)	R_{rm}	W_{rm}	N_{rm}
Total group (t)	R_{tm}	W_{tm}	N_{tm}

In this contingency table, the rows show two levels of group membership (reference group and focal group), and the columns record the frequency counts of right answers and wrong answers (omissions were included). Each item being analyzed can have M distinct 2x2 tables, with each table representing a single level of the item. For the M th score level, the data can be summarized as follows: R_{fm} and R_{rm} denote the number of examinees in the reference and focal groups, respectively, who answered correctly. In contrast, W_{fm} and W_{rm} are the numbers of examinees in the reference and focal groups who answered incorrectly (Zwick, 2012).

For the M-H procedure, the null hypothesis (H_0) and a particular alternative hypothesis (H_a), provided by Mantel and Haenszel (1959), can be expressed as:

$$H_0: \frac{\left[\frac{R_m}{W_m} \right]}{\left[\frac{R_{fm}}{W_{fm}} \right]} = 1 \quad (2)$$

$$m = 1, 2, \dots, M$$

versus

$$H_a: \frac{R_m}{W_m} = \alpha \left[\frac{R_{fm}}{W_{fm}} \right] \quad (3)$$

$$m = 1, 2, \dots, M \text{ and } \alpha \neq 1.$$

Dorans and Holland (1993) described the H_o and H_a in the following way:

In other words, (for H_0) the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and then reference group, across all M levels of the matching variable...Note that when $\alpha_m = 1$, the alternative hypothesis reduces to the null DIF hypothesis. The parameter α is called the *common odds ratio* in the M 2-by-2 tables because under H_a , the value of α is the odds ratio that is the same for all m . (p. 39–40)

The M-H chi-square statistic, which was reported as “the uniformly most powerful unbiased test of H_o versus H_a ” (Zwick & Erickson, 1989, p. 58), can be defined as follows:

$$Q_{MH} = \frac{\left[\sum_m (R_{m1} - E(R_m)) - \frac{1}{2} \right]^2}{\sum_m Var(R_m)} \quad (4)$$

where $E(R_{rm}) = \frac{N_{rm} R_{tm}}{N_{tm}}$, $Var(R_{rm}) = \frac{N_{rm} R_{tm} N_{fm} W_{tm}}{N_{tm}^2 (N_{tm} - 1)}$. In the numerator of Q_{MH} , according to

Holland and Thayer (1988), the $-\frac{1}{2}$ serves as a *continuity correction* to provide:

[The] effect of the continuity correction is to improve the calculation of the observed significance levels using the chi-square table rather than to make the size of the test equal to the nominal value. Hence, simulation studies routinely find that the actual size of a test based on [the corrected version] is *smaller* than the nominal value...The continuity correction is simply to improve the approximation of a discrete distribution ...by a continuous distribution. (p. 135)

In their original work, Mantel and Haenszel (1959) also included the estimate of the *constant odds ratio*, which can be expressed as:

$$\alpha_{MH} = \frac{\left(\frac{\sum_m R_{rm} W_{fm}}{N_{tm}} \right)}{\left(\frac{\sum_m R_{fm} W_{rm}}{N_{tm}} \right)} \quad (5)$$

α_{MH} is an effect size estimate of DIF (ranged from 0 to ∞), and it was assumed to be constant over all levels. More recently, Holland and Thayer (1985) adapted α_{MH} and then converted it from odds ratio to log odds, namely, the Mantel-Haenszel delta difference (MH D-DIF) statistic. *MH D-DIF* is defined as:

$$MH\ D-DIF = -2.35 \ln(\alpha_{MH}). \quad (6)$$

Because *MH D-DIF* conducts estimates in the ETS delta metric, researchers can easily interpret whether or not an item favors one group over another based on the value and, more importantly, on the direction of an *MH D-DIF* value. For instance, an *MH D-DIF* value of -.5 means that this

test item is estimated to be more difficult for the focal group than for the reference group by an average of 1/2 of a delta point, conditional on ability. If the value is 1 instead, then that means that this item favors the focal group.

Based upon the results obtained from the *MH D-DIF* statistic, ETS developed three categories to identify test items with varying degrees of DIF. These three categories are A (*negligible or non-significant DIF*), B (*slight to moderate DIF*), and C (*moderate to large DIF*). The classification decisions depend on two factors: (1) the absolute value of *MH D-DIF* and (2) the significance test. Table 2 describes the criteria that are utilized during the procedures of classifying a test item into three categories.

Table 2

Definition of the DIF Categories A, B, and C Based on the MH D-DIF Statistic

Category	Absolute Value and Significance of MH D-DIF
A	MH D-DIF not significantly different from 0
	OR absolute value <1.0
B	MH D-DIF significantly different from 0 and absolute value of at least 1.0
	AND EITHER
	(1) Absolute value at least 1 but less than 1.5
C	OR (2) Absolute value at least 1 but not significantly greater than 1 (.05 level)
	Absolute value of MH D-DIF at least 1.5 and significantly greater than 1

Note. Reproduced from Zieky (1993) and Longford, Holland, & Thayer (1993)

When detecting DIF in multiple groups, the M-H chi-square statistic conducts a comparison between one reference group and one focal group, which leads to an increase in the Type I error rate. Two alternative methods can be used to address this issue. The first approach is

to use a Bonferroni-adjusted alpha level to reduce the family-wise Type I error rate; however, the power of detecting DIF in multiple groups will be adversely impacted. A better method is provided in a simulation study of Penfield (2001), called a generalized Mantel-Haenszel test (GMH; Somes, 1986). In the case of assessing DIF in dichotomous items among multiple groups, Penfield (2001) showed that GMH is superior to both (i.e., the M-H chi-square statistic and M-H with a Bonferroni-adjusted alpha level) in terms of greater power and a lower Type I error rate.

More recently, Fidalgo and Scalón (2010) carried on the GMH study and evaluated its utility for assessing DIF items across multiple groups simultaneously by using simulated data. For DIF assessment, this study conducted omnibus tests by the use of “ Q contingency tables with dimensions $R \times C$ ” (p. 61) where Q , R , and C are, respectively, the levels of a matching variable, the number of groups being compared, and the levels of response variable. That is to say, in a single run, a researcher can efficiently detect whether an item exhibits DIF or not. Consistent with Penfield’s finding (2001), this study also concluded that the GMH had several advantages over pairwise methods, one being that the GMH provided a lower Type I error rate and greater power. More importantly, the GMH was extended to assess DIF items for both dichotomous and polytomous items across several groups simultaneously (Fidalgo & Scalón, 2010). If the null hypothesis is rejected when applying GMH statistics, it is necessary to employ a post hoc paired comparison in order to further investigate the groups between which DIF exhibits (Fidalgo & Scalón, 2010; Penfield, 2011).

Q_j Statistic

Based upon the foundation of Lord’s (1977, 1980) original Wald χ^2 test, researchers Kim, Cohen, and Park (1995) presented a DIF detection approach in dichotomous items for multiple-

group contexts, namely the Q_j statistic, by comparing vectors of item parameters among three or more groups simultaneously (Fidalgo & Scalton, 2010). Kim, Cohen, and Park (1995) illustrated the utility of the Q_j statistic via assessing DIF items in a data collection, described as follows. There were 14 multiple choice items with four options per item. There was a total of three groups, and each group had 200 examinees. To fit the data sets, the 2-PL model was employed, and item parameters were obtained via MML estimation. The Lord's chi-square test was included in this study to provide comparisons between a multiple-group method and a two-group method.

Like other multiple-group methods, the Q_j statistic has advantages over traditional two-group methods; a single significance test that contains all possible comparisons between a reference and a focal group is more time efficient than several pairwise comparisons. Furthermore, the Q_j statistic requires a smaller number of comparisons than two-group methods, which leads to a better controlled Type I error rate (Penfield, 2001). Unfortunately, the Q_j statistic suffers from a few limitations. First, a moderate to large sample size is essential for the Q_j statistic to perform adequately; however, in practice, it is difficult to satisfy this prerequisite, because focal groups often have a smaller sample size. Additionally, Penfield (2001) specified that

the Q_j statistic does not consider the density of examinees in the sample along the ability continuum, and thus may signal DIF in regions of the ability scale with sparse data. This constraint is known to adversely affect the performance of Lord's chi-square method (Camilli & Shepard, 1994), and likely has a similar implication for the performance of the Q_j statistic. (p. 237)

Multiple Indicator Multiple Cause (MIMIC) Models

Research has often presented multiple group DIF in the context of IRT (Thissen, Steinberg, & Gerrard, 1986; Thissen et al., 1988, 1993) or factor analysis (Jöreskog, 1971; Millsap & Yun-Tein, 2004; Muthén & Christoffersson, 1981; Woods, 2009). A popular method for the latter type is a MIMIC model. MIMIC models received more widespread attention after a series of publications by Bengt Muthén (1985, 1988, 1989). It is a popular method in which at least one observed variable (i.e., causal indicator) is used to predict a latent variable (Jöreskog & Goldberger, 1975). The figure below, presented in Woods (2009), illustrates the structure of a simple MIMIC model. Woods (2009) described Figure 1 that

[Figure 1] displays a standard unidimensional item response model (or factor model) with θ regressed on an observed grouping variable to allow for a group mean difference on theta. Item responses are regressed onto the grouping variable to test for DIF. There is evidence of DIF if group membership significantly predicts item response, controlling for any mean differences on θ . Discrimination parameters are implicitly invariant; thus this is a model of uniform DIF. (p.4)

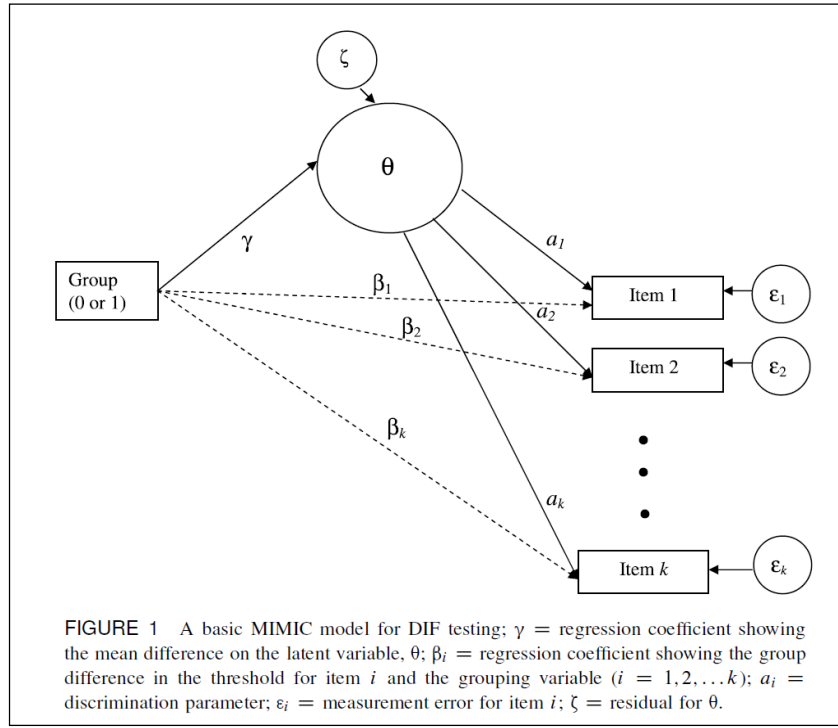


Figure 3. Basic MIMIC Model for DIF detecting

In addition, Woods and Grimm (2011) explained the procedure of assessing DIF using MIMIC model as follows:

To test for uniform DIF, item i is regressed on the latent variable (θ), and group,

$z : y_i^* = \alpha_i \theta + \beta_i z + \varepsilon_i$, where y_i^* = continuous response process that underlies a discrete y_i , α = discrimination parameter, β = regression coefficient showing the group difference in the threshold, and ε = unique factor (error). (p.342)

Woods and Grimm (2011) proposed a way by which a MIMIC model could detect non-uniform DIF by including an interaction between group and θ in the model, which is called MIMIC-interaction model. In an extensive simulation study, Woods and Grimm (2011) showed how this MIMIC-interaction model with categorical indicators can be used to test for uniform and non-uniform DIF simultaneously. MIMIC models innately have advantages in, not only

account for measurement error in the responses, as well as providing flexibility in the context of SEM such as inclusion of more than two groups, various factors, categorical or continuous covariates (Woods, 2009; Woods & Grimm, 2011). Furthermore, MIMIC-interaction model was shown to be superior to MIMIC basic model (without interaction) with respect to (1) reduction of bias while maintaining an adequate level of accuracy and (2) better performance with smaller sample size of focal groups (Woods & Grimm, 2011).

Despite the obvious advantages MIMIC-interaction models have, MIMIC models only estimate one latent variable covariance structure, making the assumption that it is equal across groups. Another disadvantage of MIMIC models is the lack of readily available software implementation of the method.

IRT Likelihood Ratio Tests

IRT likelihood ratio tests (Thissen et al., 1988), often called IRT-LR-DIF, have been reported as a more powerful and flexible method for assessing DIF than other candidate methods (Thissen et al., 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995). Tests of statistical significance using IRT-LR-DIF involve

...comparing nested two-group item response models with varying constraints to evaluate whether the response function(s) for a particular item differs for the reference and focal groups. No explicit estimation of θ is needed; θ is a random latent variable treated as missing using Bock and Aitkin's (1981) scheme for marginal maximum likelihood. The mean and variance of θ are fixed to 0 and 1 (respectively) for the reference group to identify the scale and estimated for the focal group as part of the DIF analysis. A subset of items called designated anchors are presumed invariant and used to link the metric of θ for the two groups. (Woods, 2009, p.3)

This method can be expressed as:

$$G^2(d.f.) = 2 \log \left[\frac{\text{Likelihood}[A]}{\text{Likelihood}[C]} \right], \quad (7)$$

Where Likelihood [A] and Likelihood [C] represent likelihood of the data given maximum likelihood estimates of the parameters of the model in an augmented model and a compact model, respectively. $d.f.$ is the difference between the number of parameters in the augmented model and the number of parameters in the compact model. If $G^2(d.f.)$ value is significant, then the null hypothesis of no DIF is rejected.

IRT-LR-DIF compares two models, a compact model and an augmented model (Judd & McClelland, 1989). The former is hierarchically nested within the latter. Therefore, it requires fitting the model twice in each hypothesis, which results in a much longer computation time especially when comparing more than two groups (Langer, 2008).

Improved Wald Tests

The Lord's Wald test (1977, 1980) was originally developed to compare vectors of IRT item parameters between groups:

$$\chi_i^2 = \mathbf{v}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{v}_i \quad (8)$$

Suppose this is a 2PL model, and two groups are compared with each other; then

$\mathbf{v}_i^T = [\hat{a}_{F_i} - \hat{a}_{R_i}, \hat{b}_{F_i} - \hat{b}_{R_i}]$, and, $\boldsymbol{\Sigma}_i^{-1}$ represents the estimate of the inverse of sampling variance-covariance matrix of the differences between the item parameter estimates. More specifically, the Wald statistic incorporates information from the covariance matrix of the differences of item parameter estimates between groups, along with the actual values of these differences themselves, to create a chi-square statistic. Note that the guessing parameter (c -parameter) was

not included in such tests, and Lord (1977, 1980) stated that the guess parameter should be set as equal across groups.

More recently, Langer (2008) proposed an alternative of the two-stage equating procedure for improving the original Wald's test. The equating procedure was described by Langer (2008) in the following way:

The first stage constrains the item parameters to be the same in both groups to estimate the population mean and standard deviation of the focal group relative to the reference group, assuming no DIF in any items. The second stage then treats that estimated population mean and standard deviation as fixed, and allows the item parameters to differ for the detection of DIF. (p. 11)

The final step is calculating a Wald statistic for each item, which compares reference and focal group parameters.

As stated in the introduction (Chapter I), Langer's proposed method mainly changed the linking/equating procedure and the estimation method for improving the original Wald test. With respect to the linking/equating procedure, Langer (2008) replaced the Stocking-Lord (1983) approach with concurrent calibration (Kolen & Brennan, 2004; Langer, 2008, p. 10). It appeared to be a logical change for the sake of improvement, as previous literature had shown that concurrent calibration outperforms the Stocking-Lord (1983) approach in the accuracy of estimated scores and the stability of recovering item parameters (Kim & Cohen, 2002; Petersen, Cook, & Stocking, 1983; Tian, 2011).

In addition, the improved Wald test (Langer, 2008) ought to have better performance than Lord's original approach, because the covariance matrix is estimated using the supplemented expectation maximization (SEM) algorithm (Cai, 2008; Meng & Rubin, 1991). Langer described

this algorithm to say that it “provides a convenient computational procedure for estimating the information matrix for item parameters and subsequently can provide more accurate standard errors for the estimated item parameters” (p. 12).

The advantages of the improved Wald tests are listed as follows. First, the improved Wald test is currently considered a practical alternative to IRT-LR-DIF; the latter has been shown to be the most powerful method for DIF detection (Thissen et al., 1993; Teresi et al., 2000; Wainer, 1995). Langer specified that the improved Wald test is “asymptotically equivalent to the likelihood ratio test (Thissen, et al. 1993) and less computationally-intensive” (2008, p.2). That means the procedure of employing this approach to DIF detection across several groups simultaneously would take much less computation time. Furthermore, when comparing to GMH method and basic MIMIC model, the improved Wald tests are considered superior because its utilities in detecting not only uniform DIF, but also non-uniform DIF (Woods et al., 2012). Although the improved variant of MIMIC model (i.e., MIMIC-interaction model) can detect non-uniform DIF (Woods & Grimm, 2011), the method has not yet been implemented fully and correctly in popular software programs (i.e. Mplus, R). The improved Wald tests, by contrast, have the advantage in the ease of multiple group comparisons by means of IRTPro and flexMIRT (Cai, 2012). However, this approach may suffer from some limitations. First, the Langer-improved Wald test requires more assumptions and a larger sample size compared to M-H tests. Additionally, the Wald approach theoretically could have covariates included in the model so that other factors (e.g., demographic characteristics), except theta, could be statistically controlled. However, candidate software such as IRTPro and flexMIRT currently does not allow specification of covariates.

To extend the improved Wald method proposed by Langer (2008), more recently,

researchers presented a one-stage equating procedure (Cai et al., 2011; Woods et al., 2012). These two different equating algorithms for the improved Wald test were referred to as Wald-2 and Wald-1, respectively. According to Woods et al (2012) a few statistical characteristics were shared between these two approaches:

Both Wald-1 and Wald-2 link the metric across groups simultaneously with item parameter estimation and DIF testing and should therefore improve on ad hoc linking.

Both Wald-1 and Wald-2 are implemented in IRTPro and flexMIRT and use SEM estimation for the covariance matrix. (p.5)

Woods et al. (2012) replicated and expanded upon Langer's (2008) result demonstrating the utility of Wald-1 in DIF detection among multiple groups through a simulation. Woods et al. (2012) demonstrated the superior performance of the improved Wald tests in DIF detection when sample sizes are unequal across groups. For example, reference group: focal group 1: focal group 2 = 750: 250: 250. In addition, both paper compared the improved Wald test to IRT-LR-DIF for ordinal responses (Langer, 2008; Woods et al. 2012). Table 3 shows a comparison between their simulation studies.

Table 3

Factors Used in Simulation Study

Factor	Langer (2008)	Woods, Cai, Wang (2012)
Number of groups	2 by 2 (white male, white female, black male, black female)	2, 3
Sample Size	small: 250 large: 1,000	small: 500 large: 1,000
Proportion of items with DIF	20%	25% , 50%
Number of items	5, 20, 40	24

Due to the fact that Wald-1 and Wald-2 employ different equating algorithms, it logically follows that these two approaches have distinct requirements for anchor items. Wald-1 detects DIF with designated anchor items to assess candidate items and estimate group differences. In contrast, Wald-2 did not require designated anchor items because this approach treats all items as anchor items (“test all items, anchor all items”), which was how Langer (2008) detected DIF. The results presented in Woods et al. (2012) indicate that Wald-1 outperforms Wald-2 in important metrics like correctly estimating latent mean differences and their variability. Bearing in mind that there might be a potential problem, that is, without post hoc analysis, Wald-1 lacks the ability to specify which item parameters actually differ between groups (Woods et al., 2012).

Chapter III: Methodology

The previous chapter reviewed five common methods available to detect items that may function differently (DIF) between reference and focal groups: the M-H procedure, the Q_j statistic, the MIMIC-interaction model, IRT likelihood ratio tests, and the improved Wald test (Wald-1 and Wald-2). Over the years, researchers have determined the strengths and limitations of the former four methods; however, the precise limitations of the improved Wald test (Langer, 2008; Woods et al., 2012) are still not clear due to few studies investigating this fairly new approach. This study aimed to (1) evaluate the performance of Wald-1 (test certain anchor items) with three and six groups and (2) measure the impact on test performance when anchors are inaccurately specified to varying degrees of severity.

Empirical Parameter Distributions

A simulation study was carried out using flexMIRT (Cai, 2012) to answer the research questions given in Chapter I, and repeated below.

- 1) Wald-1 requires anchors and gives results similar to those of IRT-LR-DIF. Among the comparisons between four groups, Wald-1 is more computationally efficient, expedient, and convenient than IRT-LR-DIF. How does Wald-1 perform when detecting DIF with more than two groups simultaneously?
- 2) Suppose Wald-1 uses improperly specified anchor items to detect DIF. How would this impact test performance?

Data were simulated using a 3PL model, and the item parameters were based on the 1999 Trends in International Mathematics and Science Study (TIMSS, 1999) mathematics assessment. The detailed item parameters were reported in the *TIMSS 1999 Technical Report*. A total of 110 multiple-choice mathematics items, with four options per item, from the TIMSS 1999

administration were utilized to obtain the item parameter distributions. This was accomplished by pooling item properties, creating the means and standard deviations for simulation distributions of the a , b , and c parameters. The distribution of a parameters was drawn from a uniform distribution between 0.5 and 2, the distribution of b parameters was drawn from a normal $(0, 0.7^2)$ distribution, and the distribution of c parameters was drawn from a uniform distribution between 0.1 and 0.25, parameterized thusly in order best resemble the TIMSS data.

For the 3PL model, the probability of answering an item correctly is

$$P(U_{ij}=1|\theta_i)=c_j+(1-c_j)\frac{e^{Da_j(\theta_i-b_j)}}{1+e^{Da_j(\theta_i-b_j)}},$$

where D (a scaling coefficient used to match the probit and logit metrics) equals 1.702, θ_i is examinee i 's ability, a_j is the discrimination parameter for item j , b_j is the difficulty parameter for item j , and c_j refers to the pseudo-guessing parameter for item j (Thissen & Wainer, 1982). In addition to item parameters, it is essential to plug in the ability parameter when detecting items that function differently between groups. Table 4 provides a list of the countries from which the simulation's reference and focal group ability parameters are derived. This subset of countries was selected in order to simulate small, moderate, and large differences in average ability, which is reflective of the ability heterogeneity often seen in the larger set of countries. The mean score and standard deviation for the mathematics scores of each country were reported in the *TIMSS 1999 Technical Report*.

Table 4 shows the means and standard deviations of each country in a standardized metric. Thus, the ability (θ parameter) distribution in the reference group (the U.S.) could be drawn from $\theta \sim N_r(0, 1)$. The ability parameter distribution in the focal groups (parameters were obtained by analyzing TIMSS data from Taiwan, the Netherlands, Japan, Turkey, and Jordan),

the SD and the mean post-transformation were freely estimated in the upcoming set of models, so the values of standardized means and standard deviation varied amongst focal groups.

Table 4.

Summary Statistics for Math Proficiency

Group	Country	Standardized Mean	Standardized SD
Reference	USA	0.00	1.00
Focal #1	Taiwan	0.80	1.18
Focal #2	Netherlands	0.52	0.83
Focal #3	Japan	0.96	0.91
Focal #4	Turkey	-0.85	0.98
Focal #5	Jordan	-0.72	1.17

Figure 4 below provides a visual overview of the ability parameter distributions in both the reference group and the five focal groups. The reference group is at the center.

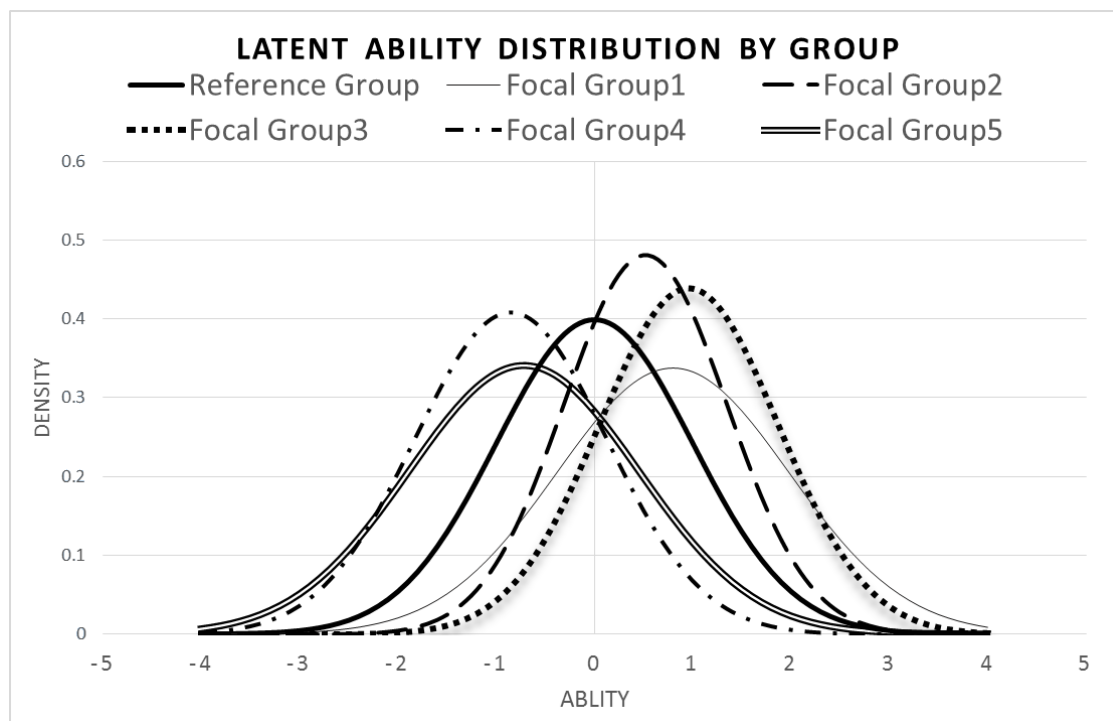


Figure 4. Ability parameters distribution of each simulated group.

Design

To evaluate the effectiveness of the Wald-1 test (which is called “test candidate items, estimate group difference with anchor items” in flexMIRT), a set of simulated data will be constructed based on the item parameters stated above. The tests consist of 50 multiple-choice items with four options. Of those 50 items, 10 are anchors. Item responses were generated using the 3PL model, with 500 replications per item per condition. The simulation in the present study consisted of 18 ($2 \times 3 \times 3$) conditions that varied in terms of (1) the number of groups (three and six groups); (2) except for the 10 anchors, the number of items that function differently on the remaining items (1, 2, and 6); and (3) the number of designated anchors that are correctly chosen (5, 8, 10). The figure below illustrates the characteristics of (2) and (3).

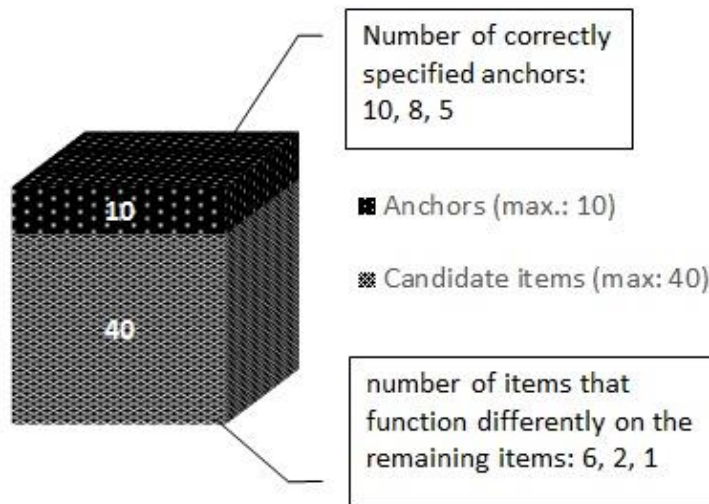


Figure 5. Simulation study: pools of anchor and candidate items.

The number of groups is three and six, with the inclusion of the reference group. Three groups were evaluated in previous studies (Langer, 2008; Woods et al., 2012); it is therefore a logical approach to match the condition that was previously evaluated (which can also serve as a

reference condition) and then investigate beyond. For the examinees in the reference group, distribution of ability parameter θ was drawn from a $\sim N(0, 1)$ distribution. An additional two or five groups with differing ability distributions (as seen in Table 4), depending on the current simulation condition, are also simulated and treated as focal groups. Table 5 (below) lists grouping conditions, as well as the proportion of simulated examinees that fall within each group. The number of participants from each country was converted to an approximate percentage, which determines the proportion of each group in the population.

Table 5

Group Conditions of the Reference and Focal Groups

Group	Country	3 groups	6 groups
Reference	USA	9,072 (50%)	9,072 (30%)
Focal #1	Taiwan	5,772 (35%)	5,772 (15%)
Focal #2	Netherlands	2,962 (15%)	2,962 (5%)
Focal #3	Japan		4,745 (15%)
Focal #4	Turkey		7,854 (20%)
Focal #5	Jordan		5,052 (15%)
Total		17,806 (100%)	35,457 (100%)

***The proportion of simulated examinees that fall within each group is based on TIMSS 1999 participating countries and students*

As stated above, 10 of the 50 items are designated as anchors. Anchors refer to a common set of items (that are DIF-free) used to place the reference group and focal groups on the same metric. Depending on the simulation condition, there can be zero, two, or five incorrectly specified anchor items. The non-anchor items will be simulated to have one, two, or six DIF items. Figure 6 illustrates the construct of three characteristics investigated through the simulated data.

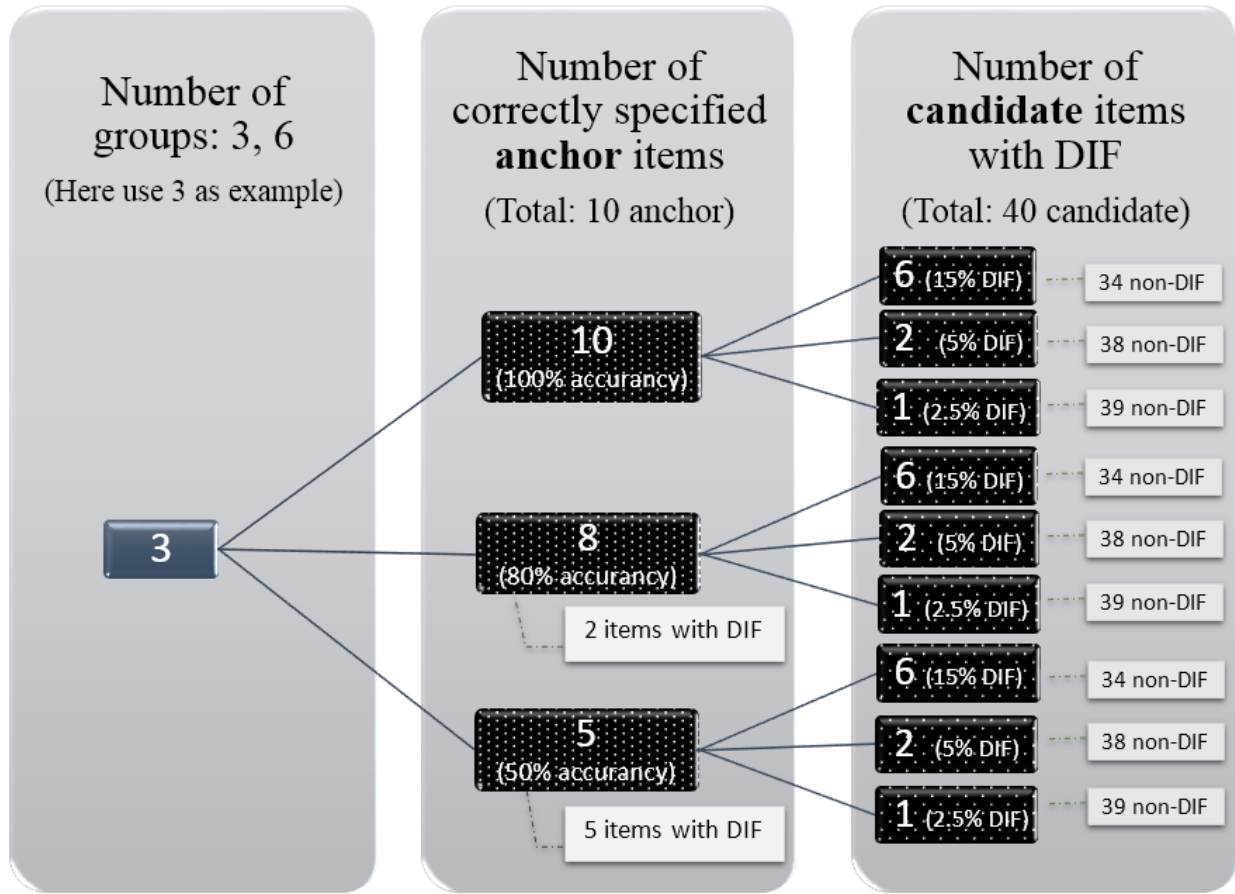


Figure 6. Illustration of the simulation design.

There were 18 simulation conditions ($2 \times 3 \times 3$). Five hundred replications were generated per condition. Among each set of 500 replications for each condition, there was a sub-condition in regard to the proportion of focal groups with DIF. That is, in three-group conditions, only one focal group had DIF; consequently, this condition had 4,500 ($1 \times 3 \times 3 \times 500$ reps) sets of generated data. In six-group conditions, either one or two focal groups had DIF. Thus, in the six-group conditions, there were 2,250 sets of generated data having one focal group that was influenced by DIF items, and the other 2,250 sets of data represented two focal groups that were influenced by DIF items. Both uniform and non-uniform DIF were used, as well as randomly selected, in

each replication in this simulation study. Uniform DIF means that one group always has advantage in answering a test item correctly at all ability level. In the case of non-uniform DIF, one group is at an advantage up to a proficiency/ability level and from that point on this group is at a disadvantage when responding to a question correctly. For the sake of simplicity in this design, the DIF direction always favored the reference group and the focal groups were disadvantaged. In terms of DIF magnitude in the uniform and non-uniform DIF conditions, if type is uniform DIF, then $b_{Reference} = b_{Focal} -$ a randomly drawn value between 0.3 and 0.7. If it is non-uniform DIF, then $a_{Reference} = a_{Focal} +$ a randomly drawn value between 0.3 and 0.7. In the present simulation, an item either differed on its a parameter or b parameter, but not both.

Outcomes

Two results were assessed: (a) statistical power and (b) Type I error rate. Statistical power was calculated as the number of studied items flagged as DIF divided by the number of non-anchor items (i.e., candidate items) with true existent DIF. For the Type I error rate, α (alpha) was calculated as the number of items flagged as DIF divided by the number of DIF-free candidate items. The results presented in the next chapter will be first described in the context of statistical power and Type I error rate, based upon certain group numbers/proportions. Subsequently, another perspective with regard to sample size and unequally sized groups will be introduced when re-evaluating the results. Additionally, a three-way Analysis of Variance (ANOVA) will be performed to find the effects of independent variables of present study.

Chapter IV: Results

The detection of items that function differently across groups is an integral part of ensuring test validity. Most of the detection methods share the limitation of being viable solely for pairwise comparisons: a limitation that is readily apparent in the evaluation of large scale, international assessments. In such circumstances, there is a vital need for comparing multiple groups simultaneously, with reasons ranging from simple time/computational efficiency to the need for properly balancing power and Type I error rate (Ellis & Kimmel, 1992; Kim, Cohen, & Park, 1995; Schmitt & Dorans, 1990). This study reviewed five common methods of detecting items that function differently between one reference group and more than one focal group. It subsequently detailed the strengths and weakness of four well-studied methods, such as the M-H procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959). Amongst available methods for evaluating DIF beyond basic binary group comparisons, the improved Wald test (Wald-1) has notably not yet been thoroughly investigated. The purpose of this study is to examine the statistical properties of Wald-1 DIF testing by means of simulated data, with the intention (1) to evaluate the performance of Wald-1 with three and six groups, and (2) to measure the impact on test performance when anchors are inaccurately specified to varying degrees of severity.

The simulation results will be discussed in terms of (a) statistical power and (b) Type I error rate. Statistical power was calculated to show the proportion of times a candidate item was correctly flagged as DIF. The Type I error rate represents the proportion of times a candidate item was falsely specified as a DIF item when it was actually DIF-free.

The Relation between Number of Groups, Anchor Contamination, and Type I Error Rate

When a pure anchor set is selected (anchor contamination is 0%), then the Type I error rate is well-controlled (i.e., constantly under 0.05) regardless of the number of groups. However,

the Type I error rates increase with a higher level of anchor contamination. Take three-group conditions, for example: when 20% of anchor items are mistakenly specified, the interquartile range of Type I error rates is 0.0375–0.1375 (median 0.075), and when anchor contamination increases to 50%, the interquartile range of Type I error rates becomes 0.225–0.3875 (median 0.325). The same increasing trend appears in the six-group conditions when observing the Type I error rates inflated at 20% contamination and 50% contamination. Figures 7 and 8 each display a box-and-whisker plot showing the three-group and six-group conditions, respectively.

Table 6

Type I Error Rate by Anchor Contamination

Type I Error Rate	Anchor Contamination					
	0%		20%		50%	
	3 Groups (rep=1500)	6 Groups (rep=1500)	3 Groups (rep=1500)	6 Groups (rep=1500)	3 Groups (rep=1500)	6 Groups (rep=1500)
Minimum	0.0000	0.0000	0.0000	0.0050	0.0000	0.0100
25%	0.0125	0.0250	0.0375	0.0450	0.2250	0.1350
Median	0.0250	0.0350	0.0750	0.0650	0.3250	0.1800
75%	0.0500	0.0450	0.1375	0.1000	0.3875	0.2550
Maximum	0.1000	0.0750	0.2875	0.1800	0.5625	0.4300

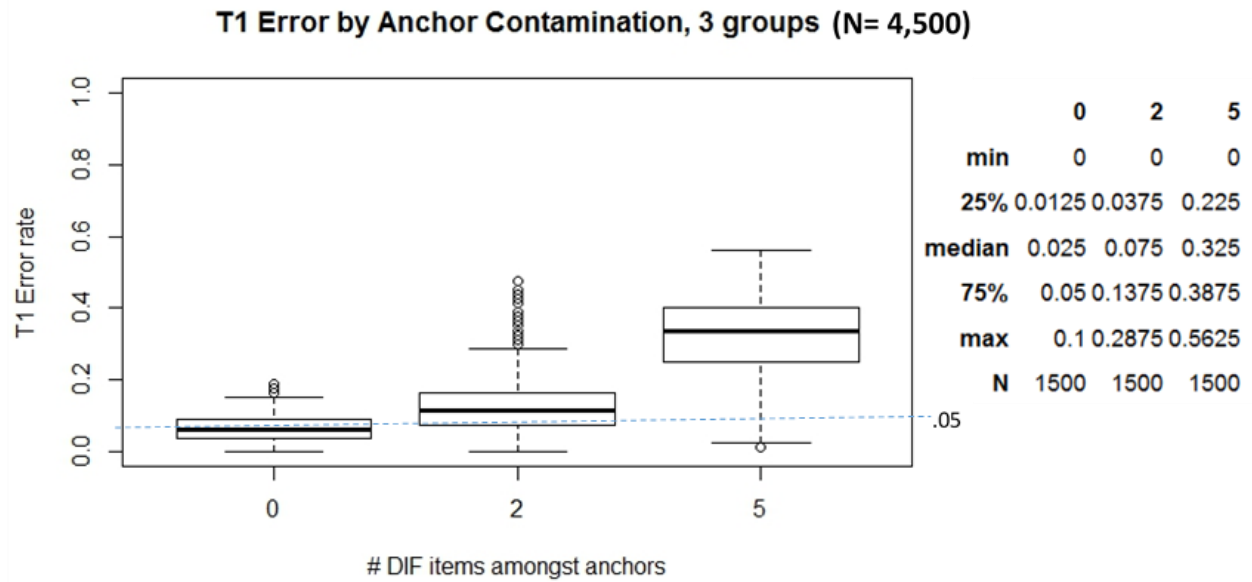


Figure 7. Box-and-whisker plot of Type I error rates by anchor contamination: three-group conditions.

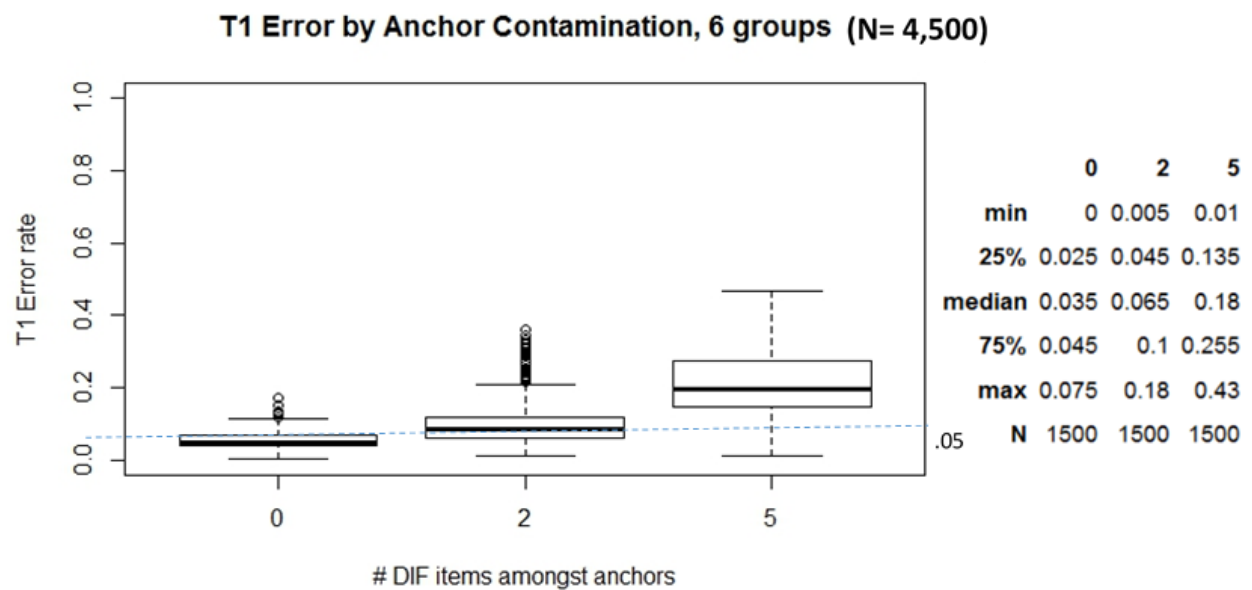


Figure 8. Box-and-whisker plot of Type I error rates by anchor contamination: six-group conditions.

Figure 9 provides a side-by-side comparison of Type I error rates of three-group versus six-group conditions. It clearly shows that three-group conditions manifest generally larger Type I error rates than those in six-group conditions at the same degree of anchor contamination (i.e., 20% and 50%). When anchor contamination is 0%, the Type I error rates of three-group and six-group conditions are well-controlled and almost equivalent.

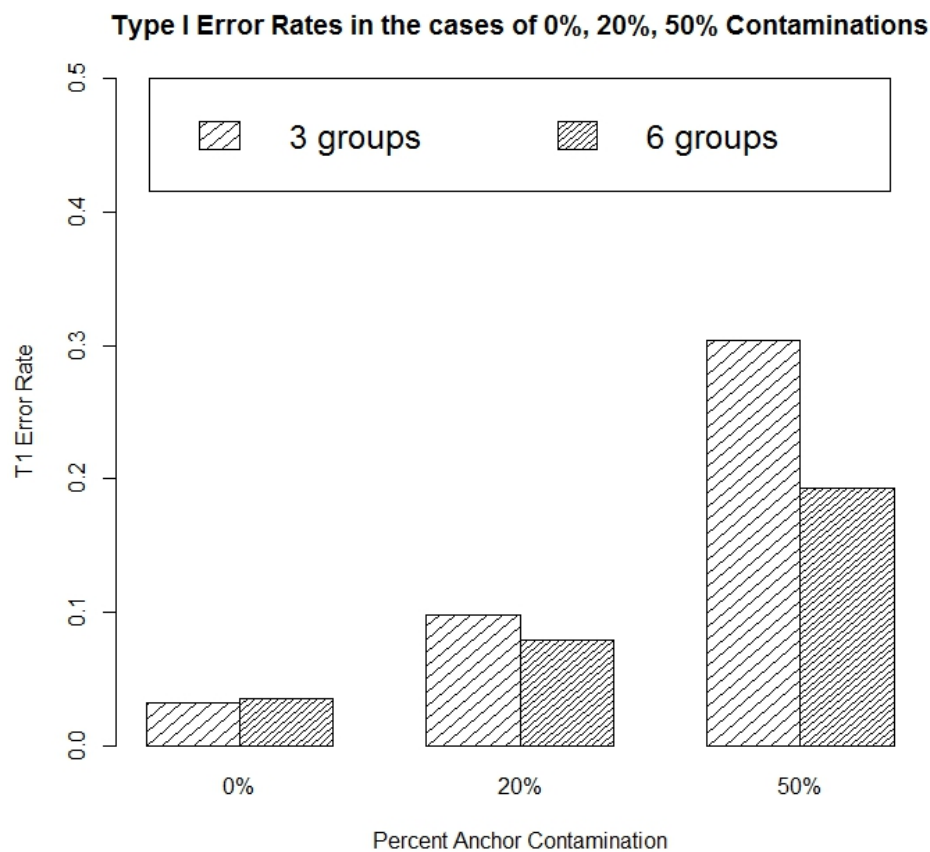


Figure 9. Side-by-side comparisons of Type I error rates for three and six groups.

As it was stated in the aforementioned Methodology section, three-group conditions could only have one focal group with DIF. However, six-group conditions could have either one or two focal groups with DIF. Figure 10 presents a comparison in Type 1 error rates between

three-group conditions and a subset of six-group conditions where only one group had DIF. The three-group conditions have Type I error rates ranging from 0–0.56 (interquartile ranges 0.0375–0.25, median 0.075), whereas the six-group conditions range from 0–0.26 (interquartile ranges 0.035–0.125, median 0.055). In short, conditions with a higher proportion of groups with DIF (collapsing across all other IVs) feature substantially higher rates of false DIF flagging.

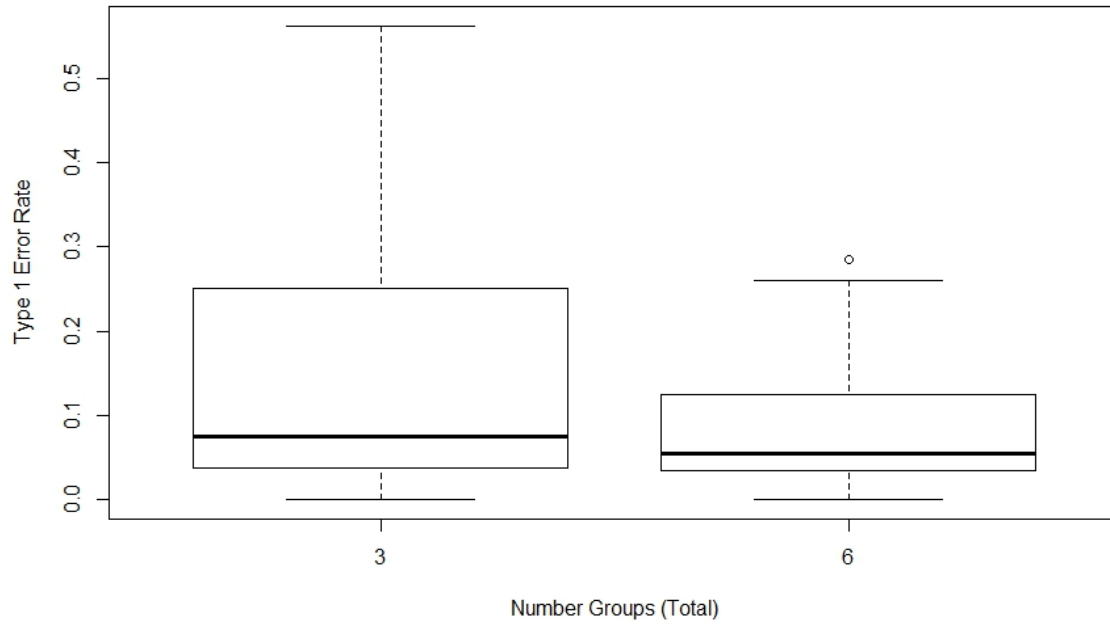


Figure 10. Type I error rates with only one DIF group: three-group versus six-group conditions.

Table 7

Summary table of Type I error rates for Figure 10

Type I Error Rate	3-Group Conditions (n=4500)	6-Group Conditions (n=2253)
Minimum	0.0000	0.0000
25%	0.0375	0.0350
Median	0.0750	0.0550
75%	0.2500	0.1250
Maximum	0.5625	0.2600

A New Factor: Proportion of Sample with DIF

Instead of rigidly adhering to how the number of groups (with DIF) impacts power and Type I error rates in DIF detection, let us assess the results from the perspective of the proportion of the sample with DIF. The simulated data had unequal sample sizes for each focal group in order to represent the realistic conditions in TIMSS. Consequently, taking the proportion of the samples with DIF into consideration seems to be a more logical approach for evaluating these results.

The effect of anchor contamination is further magnified when a larger percentage of the overall sample is in a DIF group. This is demonstrated in Figure 11 through the x-axis of the graph (i.e., percentage with DIF), with clear linear trends indicating the effect of percentage of overall sample with DIF in the presence of anchor contamination. Note that each line represents a linear regression of Type I error rates onto the percentage (of overall sample) with DIF for each anchor contamination subset.

Several key findings are revealed in Figure 11 when evaluating the effect of percentage of overall sample with DIF in the presence of varied anchor contamination levels. First, when all anchor items are correctly specified (0% contamination), then the Type I error rate is reliably under 0.05. Noticeably, regardless of the percentage of the sample that is in a DIF group, the pure anchor conditions properly controlled Type I error. Increasingly large proportions of the anchor set being contaminated are related to higher Type I error rates. Figure 11 also suggests that the anchor contamination and the percentage of total subjects with DIF can interact to greatly increase Type I error rate. Notably, when approximately 35% of overall samples are in a DIF group in the presence of 50% anchor contamination, items at this point are essentially being flagged randomly for DIF at nearly “coin-flip” rates.

In terms of statistical power, Figure 12 demonstrates that increasingly higher percentages of samples with DIF yield higher powers. Each line represents a linear regression of statistical power onto the percentage (of sample) with DIF for each anchor contamination subset. The regression line of 0% contamination shows that the power maintains at the desirable level of 0.90 and above (collapsing across all other IVs). In 20% anchor contamination conditions, it still maintains adequate statistical power above 0.8, regardless of the percentage of subjects with DIF. When 50% of anchor items are mistakenly specified, the power is consistently inferior to a smaller percentage of contamination.

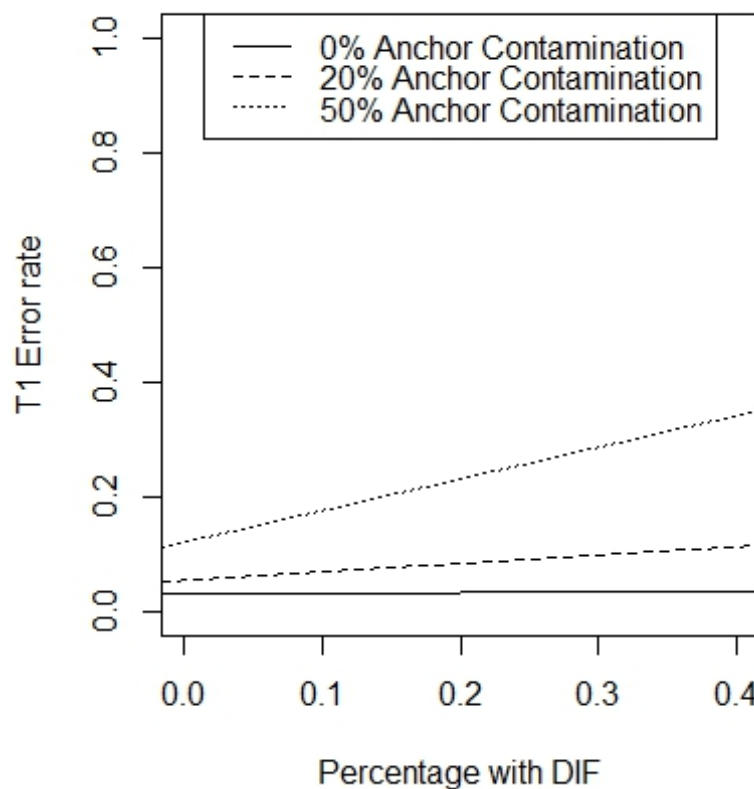


Figure 11. Type I error rates by percentage of overall sample with DIF and anchor contamination.

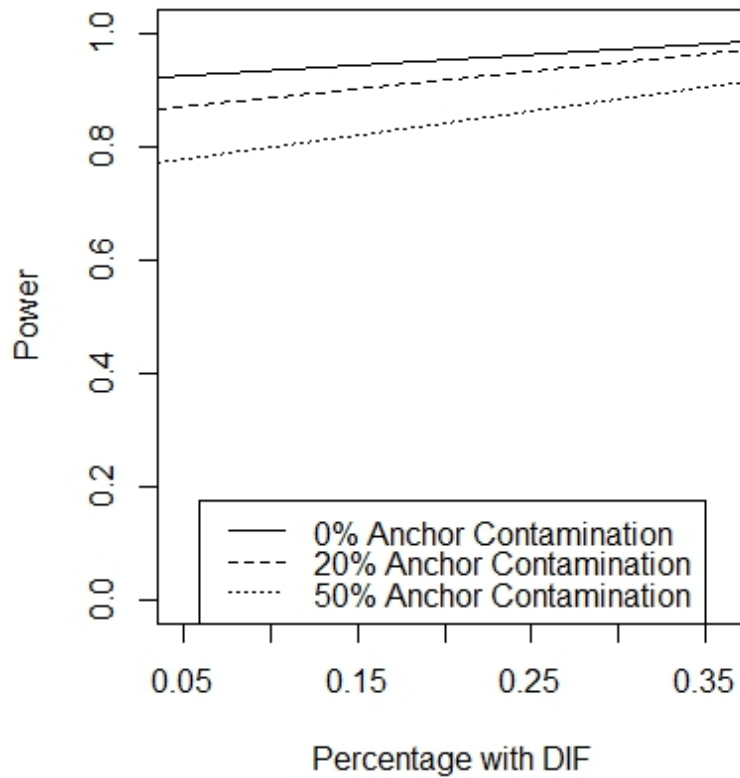


Figure 12. Statistical power by percentage of sample with DIF and anchor contamination.

Group Mean Effects

To investigate a potential group effect and its relation to the Type I error rate, further examination of Wald-1's performance was carried out. Figure 13 demonstrates that the Type I error rate for three focal groups under the condition of all other variables were controlled (i.e., 50% contamination, equal sample size, only one group with DIF in six-group conditions). The only variables that differentiate one from another are the group mean of ability (θ) and its variance.

It is shown that the high-ability groups coincide with higher false positive rates than a low-ability group when collapsing over all other variables. Notably, one high-ability group and one low-ability group share an almost equivalent magnitude of theta, but in the opposite direction (high-ability group $\theta = 0.8$, $SD = 1.18$ versus low-mean group $\theta = -0.72$, $SD = 1.17$).

However, the Type I error rate for this aforementioned high-ability group is persistently higher than the low-ability group. The results suggest that the chance of items being mistakenly flagged as DIF is inflated for high-ability groups, even after controlling for all other variables. The implications of this finding will be discussed more thoroughly in the next chapter.

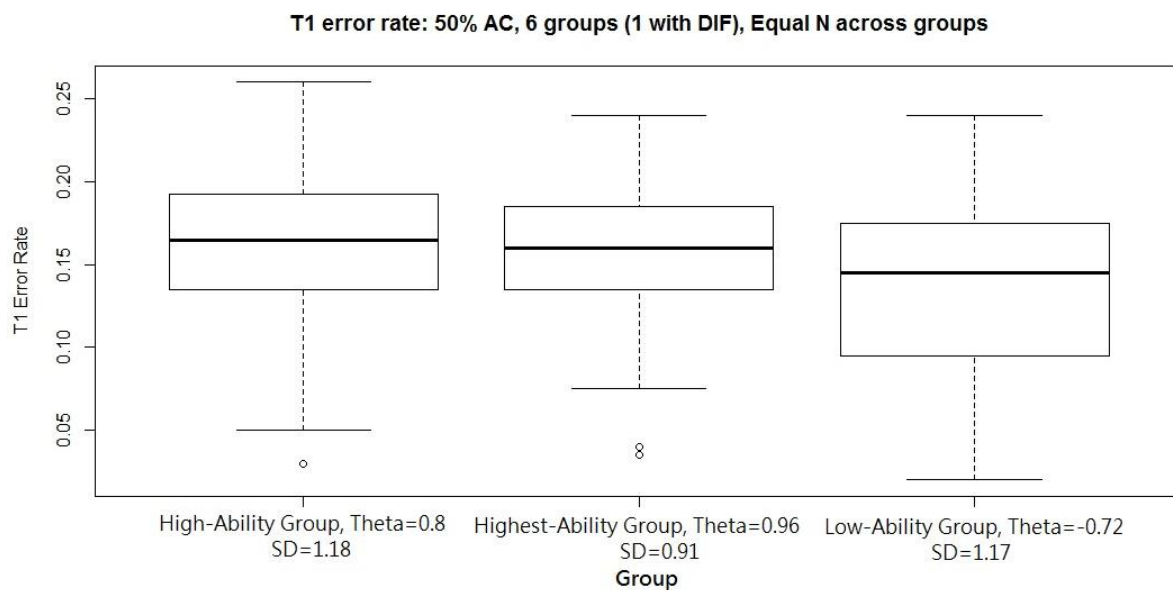


Figure 13. Type I error rate for 50% contamination, one group with DIF in six-group conditions, equal sample size.

Table 8

Summary table of Type I error rate for Figure 13.

	High-Ability Group $\theta = 0.8$ $n=136$	Highest-Ability Group $\theta = 0.96$ $n=159$	Low-Ability Group $\theta = -0.72$ $n=146$
Type I Error Rate			
Minimum	0.03	0.04	0.02
25%	0.14	0.14	0.10
Median	0.17	0.16	0.15
Mean	0.16	0.16	0.14
75%	0.19	0.19	0.18
Maximum	0.26	0.24	0.24

ANOVA Results

A three-way ANOVA was carried out to examine three characteristics: (a) anchor contamination, (b) proportion of (overall) sample with DIF, and (c) number of groups and their effects on Type I error rates. The purpose was to ensure that all major findings were thoroughly covered and discussed as needed. The three-way ANOVA results are presented in Table 9.

Table 9

Three-Way ANOVA

Main Effects and Interactions	df	Sum Square	Mean Square	F Value	Partial η^2	Pr (>F)
Percent of Respondents with DIF	5	5.080	1.020	247.225	0.121	<.000***
Anchor Contamination	2	74.820	37.410	9101.098	0.670	<.000***
Number of Groups	1	2.010	2.010	489.436	0.052	<.000***
Percent of Respondents with DIF* Anchor Contamination	10	5.270	0.530	128.300	0.125	<.000***
Percent of Respondents with DIF* Number of Groups	1	0.000	0.000	0.630	0.000	0.428
Anchor Contamination* Number of Groups	2	3.190	1.590	387.494	0.079	<.000***
Percent of Respondents with DIF* Anchor Contamination* Number of Groups	2	0.010	0.010	1.528	0.000	0.217
Residuals	8976	36.900	0.000	--	--	--

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There was no statistically significant three-way interaction between anchor contamination, proportion of overall sample with DIF, and number of groups ($p=0.217$). Thus, the focus shifts to the two significant two-way interactions. The first significant two-way interaction is the proportion of sample with DIF by anchor contamination, $F(10, 8976) = 128.3$, $p<0.001$, which indicates that the relationship between the proportion of sample with DIF and Type I error rate depends on the level of anchor contamination. Another significant (two-way)

interaction is anchor contamination by number of groups, $F(2, 8976) = 387.494$, $p < 0.001$. This shows that the relationship between anchor contamination and Type I error rate depends on the number of groups. The two-way and three-way interaction plots of means are presented in Appendix A. A simple main-effects analysis indicated that three factors (i.e., anchor contamination, number of groups, and proportion of samples with DIF) have statistically significant effects on Type I error rates. Anchor contamination showed by far the largest effect size on Type I error rates, partial $\eta^2 = 0.67$, $F(2, 8976) = 9101.098$, $p < 0.001$.

Conclusion

The results indicate that anchor contamination is overall the most influential factor in relation to Type I error rate inflation. As long as the anchor set is correctly specified, which is an inherent assumption of Wald-1, the results indicate that the Type I error rate is well-controlled (under 0.05) and the statistical power is superb (at 0.90 and above), collapsing across all other IVs. With this in mind, note that three-group and six-group conditions have almost equivalent desirable performance in the context of Type I error rates and statistical power. In short, when Wald-1 performs properly, any of the following variables is unlikely to be a significant variable in relation to the change in Type I error rate or power: (1) the number of groups (three or six) involved in the DIF detection, (2) the number of candidate items with DIF, and (3) any differences in groups' distributions of θ .

However, once the assumption of pure anchor sets is violated, the chance of the Wald-1 approach mistakenly flagging items as DIF inflates quickly, especially in the presence of another variable, the proportion of total sample with DIF. In the most severe circumstances in the simulation study (i.e., half of the anchor items were misspecified, coinciding with about 35% of overall sample with DIF), the false positive rate reaches 0.56.

Another variable, the proportion of total sample with DIF, also features substantially in the false positives in the presence of anchor contamination. The linear trends (Figure 11) suggest that a larger proportion of anchors being contaminated, along with a higher percentage of overall samples with DIF, can interact to greatly increase Type I error rates. It is reasonable to conclude that the main effect for the number of group(s) with DIF (one focal group in the three-group conditions, one or two focal groups in the six-group conditions) is an artifact of the simulation due to uneven sample sizes of each focal group as well as unbalanced group ability distributions. More specifically, three-group conditions could only have one focal group with DIF. Therefore, the possible situation is either 35% (if we chose focal group #1) or 15% (if we chose focal group #2) of the total sample being in a DIF group. Even six-group conditions could have one focal group with DIF, and the possible proportions of the total sample with DIF could be either 5%, 15%, or 20%. Put simply, it is more critical to take the proportion of the total sample with DIF into consideration rather than the number of groups with DIF when examining the variable's effect on the Type I error rate and power.

The results also reveal that the group mean effect occurred in the present simulation study. Type I error rates inflate for high-ability groups when comparing to low-ability ones, after controlling for all the other variables, such as sample size, level of anchor contamination, and number of groups with DIF. In the Chapter V, more explanations will be provided to illustrate why high-ability groups suffer inflation of the Type I error rate.

Chapter V: Discussion and Conclusions

In the pursuit of assessment validity for all examinees, it is crucial to identify items that perform differently between groups, given the same level of ability. Most DIF detection methods share the limitation of pairwise comparisons. Yet, there is a vital need for comparing multiple groups simultaneously because of accuracy through the need to decrease the Type I error rate. This study reviewed five common methods for DIF detection between reference and more than one focal groups: the M-H procedure, the Q_j statistic, the MIMIC-interaction model, IRT likelihood ratio tests, and the improved Wald test (Wald-1 and Wald-2). Researchers have determined the strengths and limitations of the former four methods; however, the precise limitations of the improved Wald test (Wald-1) are still not clear due to few studies investigating this fairly new approach. The overall purpose of this study is to examine the statistical properties of Wald-1 DIF testing by means of simulated data. This study has two specific purposes. The first is evaluating the performance of Wald-1 with three and six groups. The second is measuring the impact on test performance when anchors are inaccurately specified to varying degrees of severity. The simulation design consisted of 18 conditions that varied in terms of (a) the number of groups (three and six groups), (b) the number of candidate items with DIF, and (c) the number of designated anchors being contaminated. The simulation results were reported in the previous chapter in terms of statistical power and Type I error rate (false positives).

The findings demonstrate that Wald-1 is capable of delivering superior performance when dealing with unequal sample sizes, with uniform and non-uniform DIF exhibiting among small proportions of studied items across as many as six groups simultaneously. The most critical variable in relation to the false positives, as well as statistical power, is anchor contamination. Given that the Wald-1 approach requires a set of pure anchor items, it is logical

that the violation of this prerequisite would significantly impact its performance. The effect of anchor contamination is further magnified when a larger percentage of the overall sample is in a DIF group. The effect of percentage of overall sample with DIF (in the presence of anchor contamination) upon Type I error rate is monotonic. A 50% anchor contamination renders items essentially being flagged randomly for DIF at nearly “coin-flip” rates. However, the pure anchor conditions properly control the false positives collapsing all other variables (number of groups involved in the comparisons, number of candidate items with DIF, and number of groups with DIF).

The aforementioned observation provides some key suggestions. First, if a set of pure anchor items is specified, Wald-1 features superior performance in detecting DIF items between as many as six groups simultaneously. Outside of simulation research, psychometricians and test developers in realistic situations can hardly guarantee the purity of all specified anchor items. Therefore, using a reliable and efficient approach for anchor selection prior to implementing Wald-1 is of the utmost importance.

The results also suggest that the proportion of the sample with DIF exerts greater influence on statistical power and Type I error rates than the number of groups with DIF. With this in mind, in the application of the Wald-1 approach in empirical examples, the Type I error rate could be better controlled when the overall sample size consists of all focal groups that are much smaller than the reference group. More specifically, after controlling for the anchor contamination (if there is any), the false positives for DIF detections between several focal groups with a small sample size could be more substantially diminished than one single focal group but with a large sample size. Researchers need to be more cautious with applying the Wald-1 approach in DIF detection when handling focal groups with large sample sizes

(especially when the combined sample size for all focal groups is comparatively larger than the reference group).

Figures 7–10 show that Type I error rates in the three-group conditions are higher than those in the six-group conditions, which seems to suggest that a greater number of groups involved in the DIF detection are related to a decrease in the Type I error rate. However, further investigations demonstrate that the effect of number of groups is a by-product of this simulation design and, consequently, is not related to Wald-1's actual performance. In this case, the addition of three more focal groups in DIF detection does not necessarily decrease the Type I error rate. The decline in Type I error rates in six-group conditions merely resulted from the group mean of ability (θ) from two newly added focal groups obedient to the DIF direction in the present simulation study (i.e., the reference group is at an advantage and the focal groups are at a disadvantage). More specifically, in three-group conditions, the two focal groups both have greater group means than the reference group (0.8 and 0.52 versus 0). This means that, when anchors are contaminated and the DIF items always perform against the focal groups, the participants of focal groups will have underestimated values of theta. A contaminated anchor would reach a false conclusion that the focal groups and the reference group have equivalent ability levels. However, when there is actually a DIF-free question that is harder, the focal groups naturally perform better than the reference group due to their true group mean being higher. When comparing to what this DIF-free test item expects respondents to perform based on the conclusion drawn by the contaminated anchors, the Wald-1 method would falsely flag this item for DIF and consequently inflate the Type I error rate in three-group conditions. Figure 13 displays a clearer demonstration of the Type I error rate inflation for high-ability groups than for low-ability groups (with equivalent magnitudes) after controlling for all other variables.

In contrast, six-group conditions have 40% of focal groups (2 out of 5 focal groups) with a smaller group mean, and therefore it dilutes the chance of an item being mistakenly flagged as DIF. Put simply, the simulation results suggest that the number of groups involved in the DIF detection does not impact the Type I error rate. The difference in Type I error rate between three-group conditions and six-group conditions actually resulted from the group mean of the selected focal groups.

Conclusions

The results indicate that anchor contamination is overall the most influential factor in relation to Type I error rate inflation. As long as the anchor set is correctly specified, which is an inherent assumption of Wald-1, the results indicate that the Type I error rate is well-controlled (under 0.05) and the statistical power is superb (at 0.90 and above), collapsing across all other IVs. With this in mind, note that three-group and six-group conditions have almost equivalent desirable performance in the context of Type I error rates and statistical power. In short, when Wald-1 performs properly, any of the following variables is unlikely to be a significant variable in relation to the change in Type I error rate or power: (1) the number of groups (three or six) involved in the DIF detection, (2) the number of candidate items with DIF, and (3) any differences in groups' distributions of θ .

However, once the assumption of pure anchor sets is violated, the chance of the Wald-1 approach mistakenly flagging items as DIF inflates quickly, especially in the presence of another variable, the proportion of total sample with DIF. In the most severe circumstances in the simulation study (i.e., half of the anchor items were misspecified, coinciding with about 35% of overall sample with DIF), the false positive rate reaches 0.56.

Another variable, the proportion of total sample with DIF, also features substantially in the false positives in the presence of anchor contamination. The linear trends (Figure 11) suggest that a larger proportion of anchors being contaminated, along with a higher percentage of overall samples with DIF, can interact to greatly increase Type I error rates. It is reasonable to conclude that the main effect for the number of group(s) with DIF (one focal group in the three-group conditions, one or two focal groups in the six-group conditions) is an artifact of the simulation due to uneven sample sizes of each focal group as well as unbalanced group ability distributions. More specifically, three-group conditions could only have one focal group with DIF. Therefore, the possible situation is either 35% (if we chose focal group #1) or 15% (if we chose focal group #2) of the total sample being in a DIF group. Even six-group conditions could have one focal group with DIF, and the possible proportions of the total sample with DIF could be either 5%, 15%, or 20%. Put simply, it is more critical to take the proportion of the total sample with DIF into consideration rather than the number of groups with DIF when examining the variable's effect on the Type I error rate and power.

The results also reveal that the group mean effect occurred in the present simulation study. Type I error rates inflate for high-ability groups when comparing to low-ability ones, after controlling for all the other variables, such as sample size, level of anchor contamination, and number of groups with DIF. In the Discussion session, more explanations will be provided to illustrate why high-ability groups suffer inflation of the Type I error rate.

Limitations and Future Directions for Research

The present study investigated the statistical properties of the Wald-1 approach for assessing DIF items across multiple groups simultaneously by using simulated data. In order to examine the Wald-1 approach's performance under "realistic" conditions, the simulated data were generated by mimicking sample sizes and parameters within TIMSS. However, the selection of countries that provided parameters to construct the focal groups in the simulation did not have an overall perfectly balanced group mean. Ideally future simulation studies would be advised to construct a well-balanced theta distribution across reference and focal groups. Additionally, if one is planning to design a simulation similar to the present study, it will be worth setting out a DIF direction that goes meaningfully both ways (i.e. test items that would favor and not favor the focal groups relative to the reference group).

As this research emphasized the examination of the performance of the Wald-1 approach in various conditions of simulated environments, further investigations on other variables and their applications would be informative. Future directions include comparisons between Wald-1 and other DIF-detection approaches other than IRT-LR-DIF, an extension of the number of focal groups, and exploring efficient anchor selection methods (prior to implementing Wald-1).

References

- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, 34(4), 807–816.
- Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). *Literature and U.S. history: The instructional experience and actual knowledge of high school juniors* (NAEP Rep. No. 17-HL-01). Princeton, NJ: Educational Testing Service.
- Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.
- Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 61, 309–329.
- Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bi-factor analysis. *Psychological methods*, 16(3), 221.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.
- Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different*

- groups on the items in a test*. Princeton, NJ: Educational Testing Service.
- Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, 28(1), 61–75.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.
- Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, 77, 177–184.
- Ercikan, K. (1999, April). *Translation DIF on TIMSS*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, QC.
- Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391–407). Dordrecht, Netherlands: Kluwer.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17(3), 301–321.
- Fidalgo, Á. M., & Scalón, J. D. (2010). Using generalized Mantel-Haenszel statistics to assess DIF among multiple groups. *Journal of Psychoeducational assessment*, 28(1), 60–69.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological*

- Measurement*, 29(4), 278–295.
- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, QC.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 3847.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel- Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
- Houts, C. R., & Cai, L. (2012). flexMIRT TM: Flexible multilevel item factor analysis and test scoring user's manual version 1.0. Seattle, WA: Vector Psychometric Group.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631–639.
- Jöreskog, K. G., & Sörbom, D. (2007). *LISREL 8.80*. Chicago, IL: Scientific Software International.
- Judd, C. M., & McClelland G. H. (1989). *Data analysis: A model-comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.
- Kaplan, D. (2004). *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage.

- Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26(1), 25–41.
- Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261–276.
- Kolen, M. J., & Brennan, R. L. (2004). Test equating, linking, and scaling: Methods and practices.
- Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.
- Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F. M. (1952). A theory of test scores (Psychometric Monograph No. 7). *Iowa City, IA: Psychometric Society.*
- Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, Netherlands: Swets and Zeitlinger.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2), 127–143.
- Meng, X., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899–909.
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479–515.
- Muthe'n, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10, 121–132.
- Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Erlbaum.
- Muthe'n, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557–585.
- Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46, 407–419.
- Ong, S. L. (2007). Comparing two language version of science achievement tests using differential item functioning. *Malaysian Journal of Educators and Education*, 22, 45–59.
- Pei, L. K., & Li, J. (2010). Effects of Unequal Ability Variances on the Performance of Logistic Regression, Mantel-Haenszel, SIBTEST IRT, and IRT Likelihood Ratio for DIF Detection. *Applied Psychological Measurement*, 34(6), 453–456.
- Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, 14, 235–259.

- Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8, 137–156.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8(2), 164.
- Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, 27, 67–81.
- Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159–194.
- Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *American Statistician*, 40, 106–108.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important? *Journal of Applied Psychology*, 89(3), 497.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27, 361–370.
- Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, 19, 1651–1683.

- Thissen, D. (1987). Discussant comments on the National Council on Measurement in Education symposium, unexpected differential item performance and its assessment among black, Asian-American, and Hispanic students. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (Research Memorandum No. 87-1) (pp. 1–6). Princeton, NJ: Educational Testing Service.
- Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47(4), 397–412.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118–128.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item-response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT* (Unpublished doctoral dissertation). Boston College, Chestnut Hill, MA.
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, 8, 157–186.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number

- of observations is larger. *Transactions of the American Mathematical Society*, 54, 426–482.
- Woods, C. M. (2009). Testing for differential item functioning with measures of partial association. *Applied Psychological Measurement*, 33(7), 538–554.
- Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement*, 35(2), 145–164.
- Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF Testing With Multiple Groups Evaluation and Comparison to Two-Group IRT. *Educational and Psychological Measurement*, 73(3), 532–547.
- Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35, 339–361.
- Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.
- Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. Retrieve from <http://origin-www.ets.org/Media/Research/pdf/RR-12-08.pdf>
- Zwick, R., Donoghue, J., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.
- Zwick, R., & Ercikan, K. (2005). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26(1), 55–66.

Appendix A: Three-Way ANOVA Results

A three-way ANOVA was performed to further examine three characteristics: (a) anchor contamination, (b) proportion of (overall) sample with DIF, and (c) number of groups and their effects on Type I error rates.

This appendix provides the interaction plots of (group) means in Table 9, including two-way and three-way interactions (between anchor contamination, proportion of overall sample with DIF, and number of groups). Figure 14 shows that the first significant two-way interaction is the proportion of samples with DIF by anchor contamination, $F(10, 8976) = 128.3, p < 0.001$.

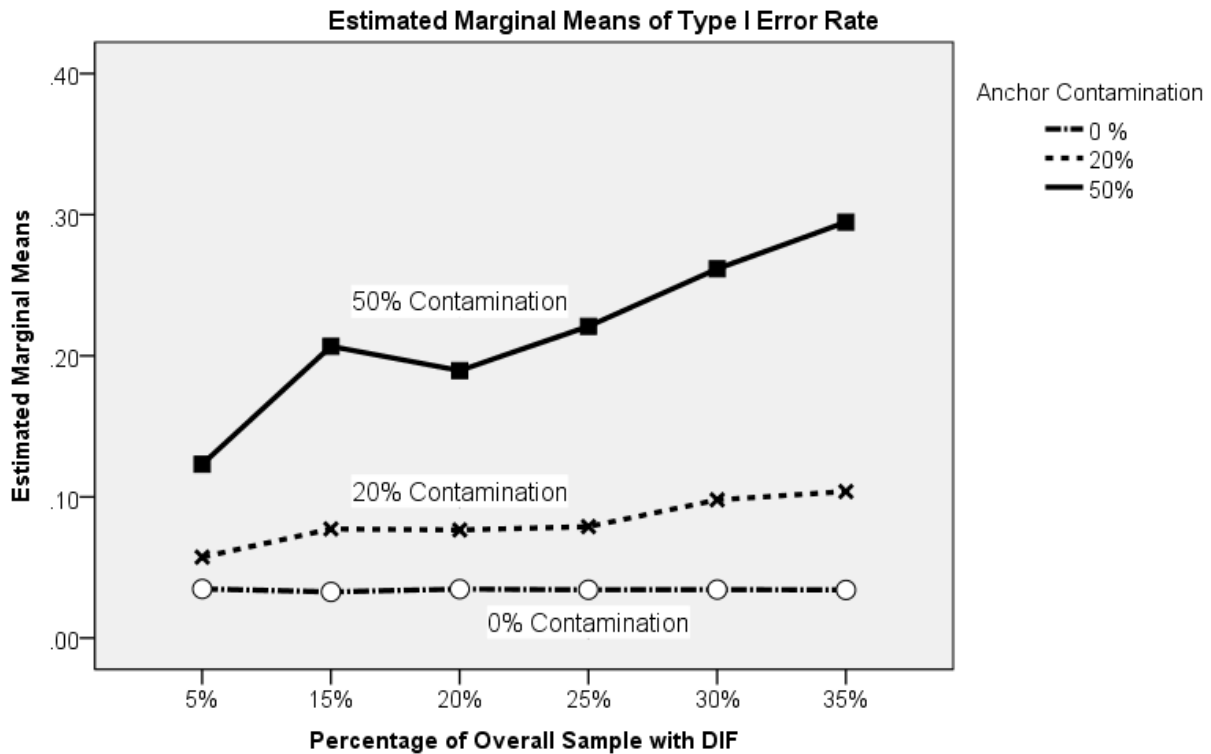


Figure 14. Interaction plot: percentage of samples with DIF by anchor contamination.

Another significant two-way interaction is anchor contamination by number of groups, $F(2, 8976) = 387.494, p < 0.001$ (Figure 15).

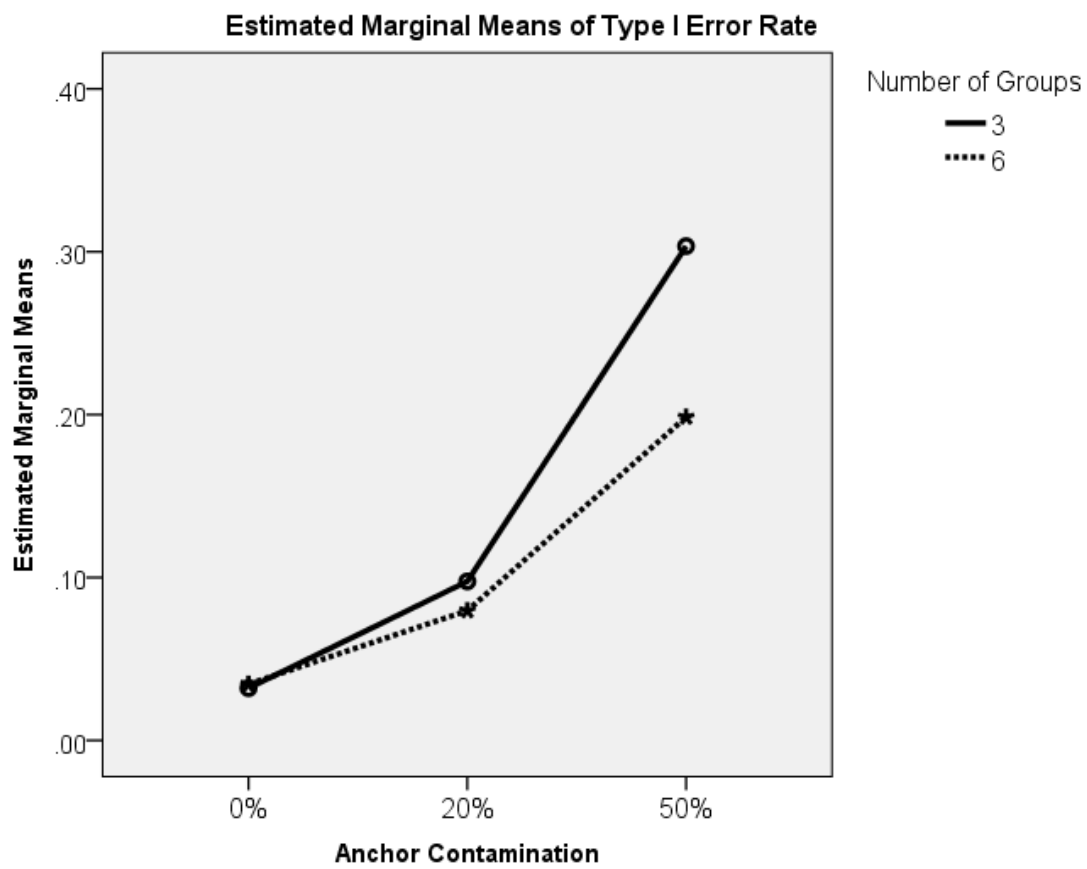


Figure 15. Interaction plot: anchor contamination by number of groups.

The last two-way interaction (percentage of overall sample with DIF by number of groups) is not statistically significant (Figure 16).

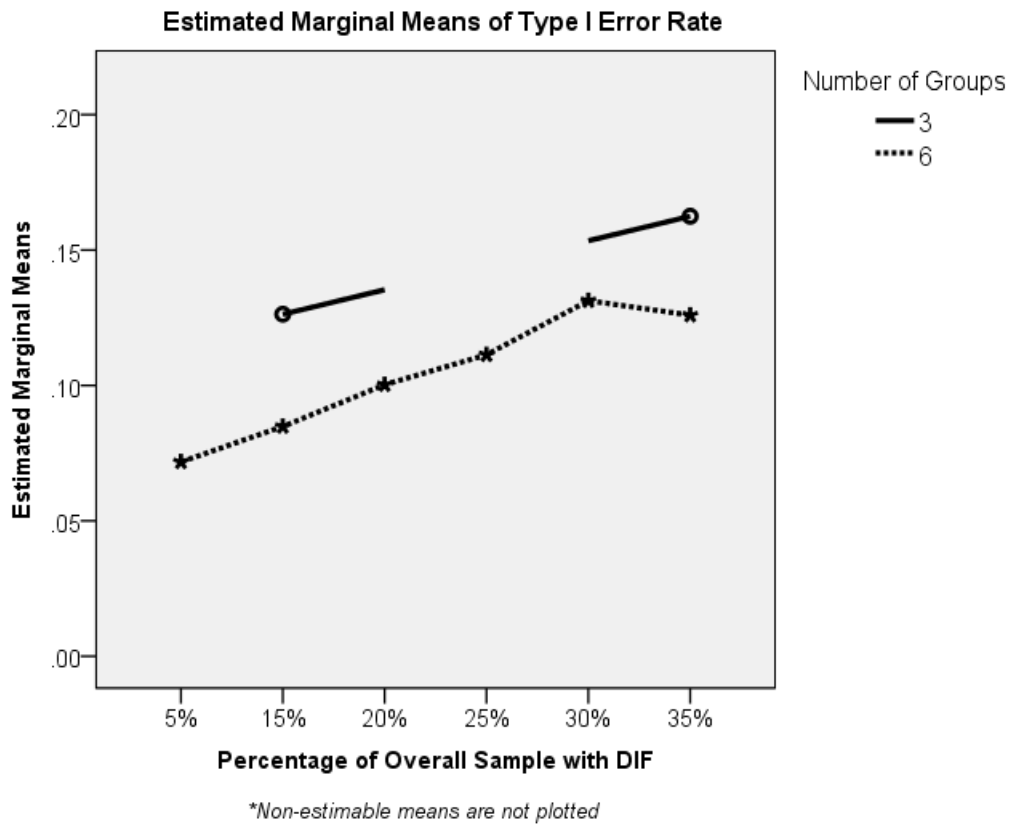


Figure 16. Interaction plot: percentage of the overall sample with DIF by number of groups.

There is no statistically significant three-way interaction between anchor contamination, proportion of overall sample with DIF, and number of groups ($p=0.217$). Figures 17 and 18 represent a three-way interaction in three-group and six-group conditions, respectively.

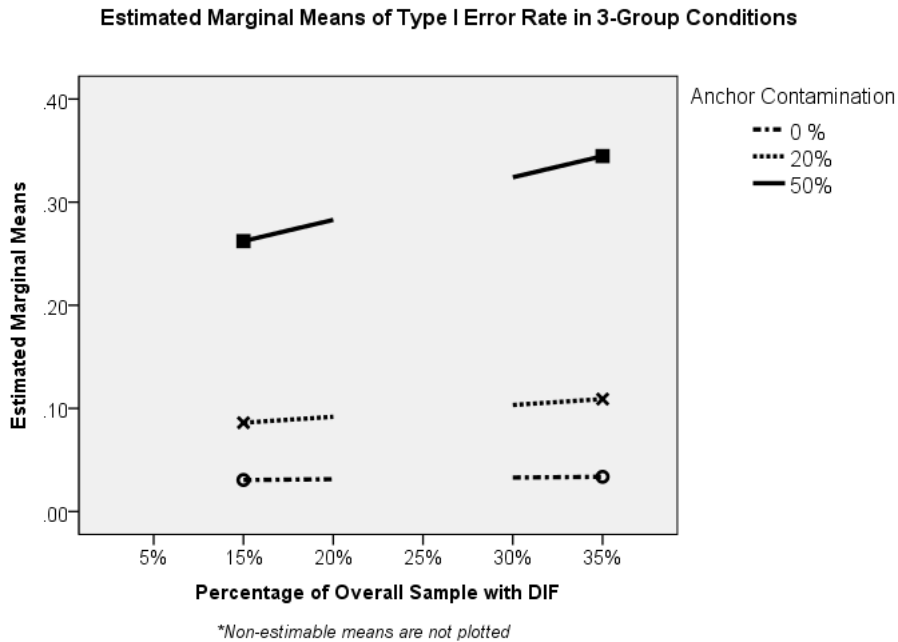


Figure 17. Three-way interaction plot (three-group conditions)

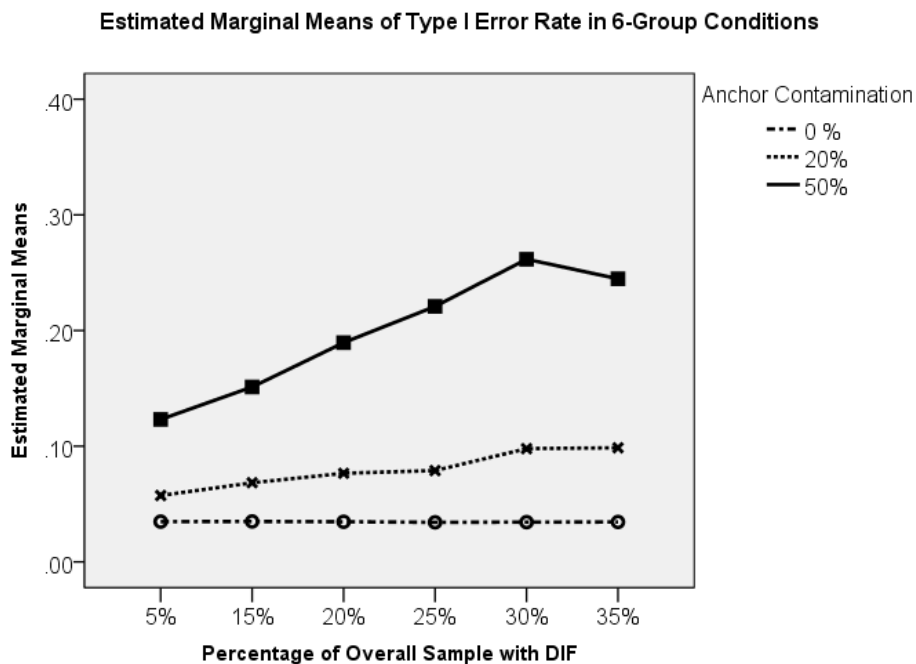


Figure 18. Three-way interaction plot (six-group conditions).